

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - TP.HCM  
KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

**Lớp: CQ2022/1 - Nhóm: CQ.BI#03**

**Thành viên nhóm:**

**22120091**

**Phạm Khánh Hân**

**22120233**

**Nguyễn Thị Tú Ngọc**

**22120258**

**Quách Quỳnh Như**

**22120284**

**Dương Kim Phụng**

Hồ Chí Minh, ngày 31 tháng 12 năm 2025

## MỤC LỤC

<b>I. Mô tả nguồn dữ liệu .....</b>	<b>8</b>
1.    File airlines.csv (Danh mục hãng hàng không) .....	8
2.    File airports.csv (Danh mục sân bay).....	8
<b>II. Thiết kế NDS .....</b>	<b>13</b>
1.    Thiết kế và đánh giá dạng chuẩn (Normalization Analysis) .....	13
2.    Mô tả NDS.....	14
a.    NDS_Airline.....	14
b.    NDS_Airport.....	15
c.    NDS_Reason .....	15
d.    NDS_Distance.....	15
e.    NDS_Time .....	16
f.    NDS_Flight .....	16
<b>III. Thiết kế DDS .....</b>	<b>18</b>
1.    Thiết kế.....	19
2.    Mô tả DDS.....	20
a.    Dim_Date .....	20
b.    Dim_Time_Of_Date .....	20
c.    Dim_Airline .....	21
d.    Dim_Airport .....	21
e.    Dim_Reason .....	22
f.    Fact_Flight.....	23
<b>IV. Thiết kế Metadata .....</b>	<b>24</b>
1.    ds_table_type .....	24
2.    ds_table.....	24
3.    ds_column .....	25
4.    status_table .....	25
5.    data_flow.....	26
6.    dq_rule_category.....	27

7. dq_rule_action .....	27
8. dw_user .....	27
9. dq_rule .....	27
10. dq_notification.....	28
11. usage_log .....	29
<b>V. Business rules .....</b>	<b>29</b>
1. NDS_Reason .....	29
2. NDS_Distance .....	29
3. NDS_Time.....	30
4. NDS_Flight .....	30
5. Dimensions .....	30
6. Fact_Flight.....	31
<b>VI. Mapping giữa các tầng dữ liệu.....</b>	<b>31</b>
<b>Source → Stage → NDS.....</b>	<b>31</b>
1. Bảng mapping.....	31
1.1. Mapping Airline .....	31
1.2. Mapping Airport .....	32
1.3. Mapping Time .....	32
1.4. Mapping Reason.....	33
1.5. Mapping AirportCodeMapping .....	33
1.6. Mapping Flight .....	34
1.7. Mapping Distance .....	36
2. Phân tích Control và ETL Component.....	37
2.1. Source -> Stage .....	37
2.2. Stage -> NDS.....	47
<b>NDS -&gt; DDS.....</b>	<b>68</b>
1. Bảng Mapping .....	68
1.1. Mapping NDS_Airline → Dim_Airline (SCD Type 2) .....	68
1.2. Mapping NDS_Airport → Dim_Airport (SCD Type 2) .....	69

1.3. Mapping NDS_Reason → Dim_Reason (SCD Type 1) .....	70
1.4. Mapping NDS_Time → Dim_Date.....	70
1.5. Mapping NDS_Flight → Dim_Time_Of_Date .....	72
1.5.1. Mapping NDS_Flight → Fact_Flight .....	72
2. Giải thích các Control và ETL Component.....	74
<b>VII. Tự động hóa ETL .....</b>	<b>91</b>
1. Giải pháp.....	91
2. Quy trình.....	92
3. Kết quả .....	93
<b>VIII. Olap Cube.....</b>	<b>94</b>
1. Mục tiêu.....	94
2. Data source view (DSV).....	94
3. OLAP cube.....	96
4. Phân cấp chiều.....	99
a. Dim_TimeOfDay .....	99
b. DimDate .....	100
c. DimAirport .....	102
d. DimAirline .....	103
e. DimReason .....	104
<b>IX. MDX.....</b>	<b>105</b>
1. Truy vấn tổng số chuyến bay theo tháng, quý, năm .....	105
2. Top 5 hãng hàng không có nhiều chuyến bay nhất .....	107
3. Tỉ lệ chuyến bay đúng giờ (OTP) theo hãng hàng không.....	108
4. Tỉ lệ hủy chuyến theo nguyên nhân.....	109
5. Trung bình thời gian delay theo sân bay đi/đến.....	110
<b>X. Excel report.....</b>	<b>112</b>
1. Flight summary .....	112
a. Tổng quan theo mùa .....	112
b. Phân tích chi tiết theo mùa.....	113

c.	<b>So sánh các chỉ số .....</b>	113
d.	<b>Insight.....</b>	114
2.	<b>Airline report.....</b>	114
a.	<b>Tổng quan hiệu suất theo hãng .....</b>	114
b.	<b>Phân tích chi tiết theo từng hãng.....</b>	114
c.	<b>So sánh các chỉ số chính.....</b>	115
d.	<b>Insight chính .....</b>	115
e.	<b>Liên hệ với mô hình dự đoán (Is_Delayed).....</b>	115
3.	<b>Airport report.....</b>	116
a.	<b>Phân tích hiệu suất theo quy mô sân bay.....</b>	116
b.	<b>Phân tích theo sân bay nhỏ / khu vực.....</b>	116
c.	<b>Mối quan hệ giữa lưu lượng và độ trễ.....</b>	116
4.	<b>Root cause analysis.....</b>	117
a.	<b>Phân tích nguyên nhân hủy chuyến .....</b>	117
b.	<b>Phân tích nguyên nhân trễ chuyến .....</b>	117
5.	<b>Time analysis .....</b>	118
a.	<b>Tổng quan theo thời điểm trong ngày .....</b>	118
b.	<b>Phân tích chi tiết theo từng khung giờ .....</b>	118
c.	<b>So sánh các chỉ số theo khung giờ.....</b>	119
d.	<b>Insight.....</b>	119
XI.	<b>Dashboard .....</b>	120
1.	<b>Summary.....</b>	120
1.1.	<b>Tổng quan .....</b>	121
1.2.	<b>Phân tích chi tiết.....</b>	121
1.2.1.	<b>Hiệu suất của các hãng hàng không .....</b>	121
1.2.2.	<b>Top 5 hãng hàng không có OTP cao nhất.....</b>	122
1.2.3.	<b>Tỉ lệ delay theo sân bay.....</b>	122
1.2.4.	<b>Xu hướng số chuyến bay theo tháng .....</b>	123
2.	<b>Airline: .....</b>	124

<b>2.1. Tổng quan .....</b>	124
<b>2.2. Phân tích chi tiết.....</b>	124
<b>2.2.1. Tổng số chuyến bay bị delay .....</b>	124
<b>2.2.2. Hiệu suất các hãng theo quy mô và tỉ lệ đúng giờ (OTP).....</b>	125
<b>2.2.3. Tổng nguyên nhân hủy chuyến theo từng hãng .....</b>	125
<b>2.2.4. Số chuyến bay đúng giờ theo từng hãng .....</b>	126
<b>2.2.5. Tổng số chuyến bị hủy theo hãng .....</b>	126
<b>3. Airport: .....</b>	127
<b>3.1. Tổng quan: .....</b>	127
<b>3.2. Phân tích chi tiết: .....</b>	128
<b>3.2.1. Tổng quan hiệu suất Delay &amp; Cancellation theo sân bay .....</b>	128
<b>3.2.2. Phân bố thời gian delay trung bình theo sân bay .....</b>	128
<b>3.2.3. Phân bố thời gian delay trung bình theo sân bay .....</b>	129
<b>3.2.4. Phân tích nguyên nhân hủy chuyến bay .....</b>	130
<b>3.2.5. Phân tích chi tiết nguyên nhân delay theo từng sân bay .....</b>	130
<b>3.2.6. Xếp hạng sân bay theo tỉ lệ hủy chuyến .....</b>	131
<b>3.3. Cơ chế lọc và tương tác dashboard: .....</b>	131
<b>4. Time-delay: .....</b>	132
<b>4.1. Tổng quan .....</b>	133
<b>4.2. Phân tích xu hướng delay và nguyên nhân delay theo mùa, tháng.....</b>	134
<b>4.2.1. Phân tích xu hướng, tính mùa vụ và nguyên nhân .....</b>	134
<b>4.2.2. Phân tích nguyên nhân cụ thể .....</b>	136
<b>4.2.3. Đề xuất.....</b>	136
<b>4.3. Phân tích xu hướng delay theo tuần.....</b>	137
<b>4.3.1. Tổng quan xu hướng .....</b>	137
<b>4.3.2. Phân tích nhịp độ vận hành tuần.....</b>	137
<b>4.3.3. Phân tích tác động mùa vụ lên cấu trúc tuần .....</b>	138
<b>4.3.4. Kiến nghị quản trị .....</b>	138
<b>4.4. Phân tích hiệu suất vận hành theo hãng hàng không .....</b>	139

5.	<b>Time of Day – Delay</b> .....	140
5.1.	<b>Tổng quan</b> .....	140
5.2.	<b>Tổng quan hiệu suất Delay theo khung giờ trong ngày</b> .....	140
5.3.	<b>Phân tích Delay theo Time of Day &amp; Hour (Heatmap)</b> .....	141
5.4.	<b>Phân tích xu hướng số chuyến Delay theo thời điểm trong ngày</b> .....	142
<b>XII.</b>	<b>Data Mining</b> .....	142
1.	<b>Tổng quan bài toán</b> .....	142
1.1.	<b>Mục tiêu bài toán</b> .....	142
1.2.	<b>Định nghĩa “chuyến bay trễ”</b> .....	142
2.	<b>Khám phá dữ liệu (Exploratory Data Analysis - EDA)</b> .....	142
2.1.	<b>Bài toán Mất cân bằng (Class Imbalance)</b> .....	142
2.2.	<b>Phân tích correlation matrix</b> .....	143
2.3.	<b>Feature Engineering</b> .....	143
3.	<b>Thuật toán XGBoost</b> .....	144
3.1.	<b>Giới thiệu thuật toán</b> .....	144
3.2.	<b>So sánh các thuật toán</b> .....	144
3.3.	<b>Lý do chọn XGBoost</b> .....	145
4.	<b>Chuẩn bị dữ liệu và Huấn luyện mô hình</b> .....	146
4.1.	<b>Chiến thuật chống rò rỉ dữ liệu (Data leakage)</b> .....	146
4.2.	<b>Chiến thuật xử lý mất cân bằng dữ liệu (Class Imbalance)</b> .....	146
4.3.	<b>Tối ưu hóa tham số (Hyperparameter Tuning)</b> .....	146
5.	<b>Đánh giá mô hình</b> .....	147
<b>XIII.</b>	<b>Áp dụng AI</b> .....	147
1.	<b>PowerBI Smart Narrative</b> .....	147
2.	<b>ChatPowerBI</b> .....	149
	<b>Tham khảo</b> .....	151

## I. Mô tả nguồn dữ liệu

### 1. File airlines.csv (Danh mục hãng hàng không)

Chứa thông tin về các hãng hàng không, chủ yếu được sử dụng làm bảng tra cứu để chuyển đổi mã ngắn thành tên đầy đủ.

Tên cột	Ý nghĩa nghiệp vụ	Giá trị nghiệp vụ
<b>IATA_CODE</b>	Mã IATA của hãng hàng không (2 ký tự).	Là mã định danh duy nhất của hãng hàng không.
<b>AIRLINE</b>	Tên đầy đủ của hãng hàng không.	Cung cấp thông tin hiển thị thân thiện với người dùng.

### 2. File airports.csv (Danh mục sân bay)

Chứa thông tin chi tiết về các sân bay, bao gồm định danh và thông tin địa lý, rất quan trọng cho việc phân tích không gian.

Tên cột	Ý nghĩa nghiệp vụ	Giá trị nghiệp vụ	Miền giá trị
<b>IATA_CODE</b>	Mã IATA của sân bay (3 ký tự).	Là mã định danh duy nhất của sân bay. Được sử dụng để xác định sân bay Khởi hành và sân bay Đến trong dữ liệu chuyến bay.	-
<b>AIRPORT</b>	Tên đầy đủ của sân bay.	Tên hiển thị chi tiết cho người dùng.	-
<b>CITY</b>	Thành phố mà sân bay phục vụ.	Dùng để nhóm, lọc dữ liệu theo thành phố, hoặc hiển thị vị trí trên giao diện người dùng.	-
<b>STATE</b>	Bang/Vùng mà sân bay tọa lạc.	Cung cấp ngữ cảnh địa lý rộng hơn CITY.	-

<b>COUNTRY</b>	Quốc gia mà sân bay tọa lạc.	Dùng để phân tích chuyến bay quốc tế.	USA
<b>LATITUDE</b>	Vĩ độ địa lý của sân bay.	Cùng với LONGITUDE, cho phép tính toán khoảng cách thực tế giữa hai sân bay, xác định tuyến bay tối ưu, hoặc hiển thị chính xác vị trí trên bản đồ.	Từ 13.5 đến 71.3
<b>LONGITUDE</b>	Kinh độ địa lý của sân bay.	Cần thiết cho các ứng dụng GIS (Hệ thống thông tin địa lý) và phân tích định tuyến bay.	Từ -177 đến -64.8

### 3. File flights\_batch.csv

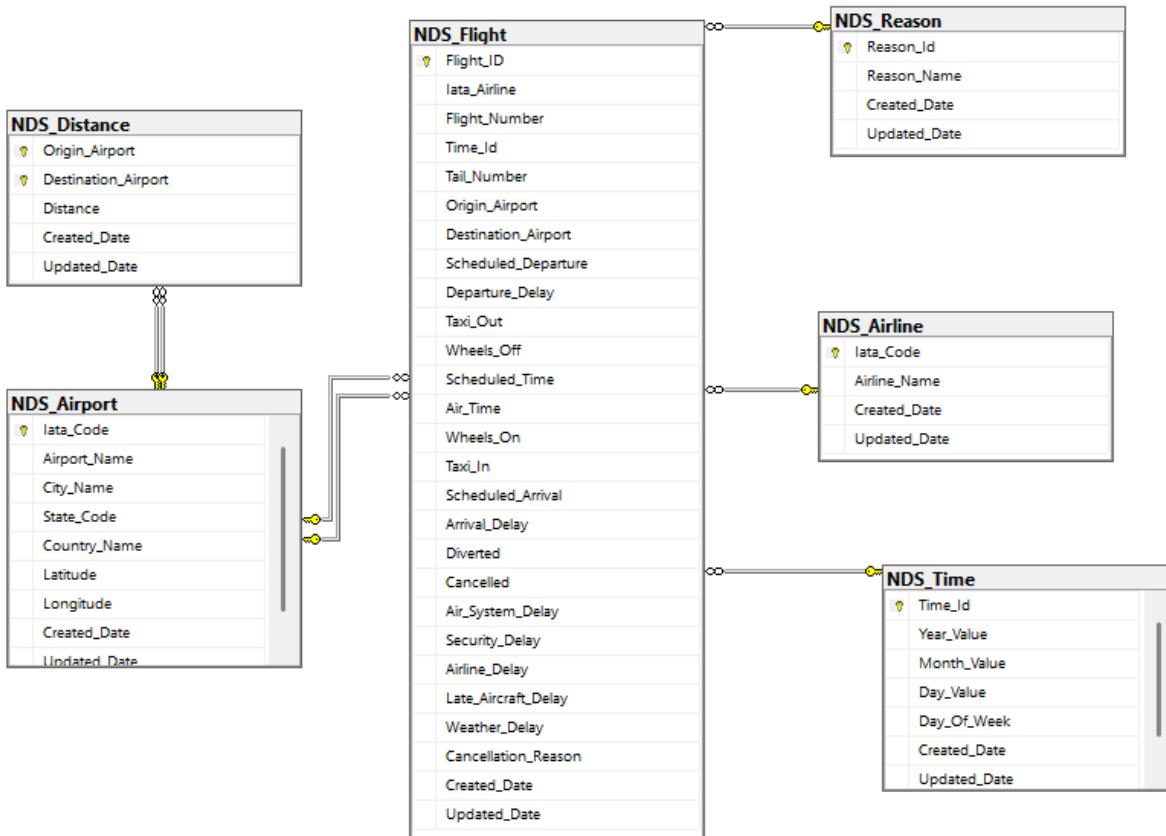
Thuộc tính	Ý nghĩa nghiệp vụ	Giải thích	Miền giá trị
<b>YEAR</b>	Năm chuyến đi	Năm của chuyến bay	2015
<b>MONTH</b>	Tháng chuyến đi	Tháng của chuyến bay	Từ 1 đến 12
<b>DAY</b>	Ngày chuyến đi	Ngày trong tháng của chuyến bay	Từ 1 đến 31
<b>DAY_OF_WEEK</b>	Ngày trong tuần	Ngày trong tuần của chuyến bay (1=Thứ Hai, 7=Chủ Nhật)	Từ 1 đến 7
<b>AIRLINE</b>	Mã hãng hàng không	Mã định danh của hãng hàng không	-
<b>FLIGHT_NUMBER</b>	Số hiệu chuyến bay	Mã định danh của chuyến bay	Từ 1 đến 9855
<b>TAIL_NUMBER</b>	Mã số máy bay	Mã định danh của máy bay	-

<b>ORIGIN_AIRPORT</b>	Sân bay khởi hành	Sân bay bắt đầu chuyến bay	-
<b>DESTINATION_AIRPORT</b>	Sân bay đến	Sân bay đến của chuyến bay	-
<b>SCHEDULED_DEPARTURE</b>	Thời điểm khởi hành dự kiến	Thời điểm cất cánh đã lên kế hoạch (được format theo dạng 24 giờ)	Từ 1 đến 2359
<b>DEPARTURE_TIME</b>	Thời điểm khởi hành thực tế	Thời điểm máy bay lăn bánh ra khỏi cổng (WHEEL_OFF) trừ đi thời gian lăn bánh ra (TAXI_OUT) (Thời gian thực tế rời cổng)	Từ 1 đến 2400
<b>DEPARTURE_DELAY</b>	Thời gian trễ khởi hành	Tổng thời gian trễ lúc khởi hành (bằng Departure_time – Scheduled_Departure) và được tính bằng phút	Từ -82 đến 1988
<b>TAXI_OUT</b>	Thời gian lăn bánh ra	Khoảng thời gian từ khi rời cổng sân bay khởi hành đến khi bánh máy bay rời mặt đất (Wheels Off) (được tính bằng phút)	Từ 1 đến 225
<b>WHEELS_OFF</b>	Thời điểm bánh rời mặt đất	Thời điểm bánh máy bay rời khỏi mặt đất (cất cánh)	Từ 1 đến 2400
<b>SCHEDULED_TIME</b>	Thời gian bay dự kiến	Tổng thời gian dự kiến cần thiết cho chuyến bay được tính bằng phút	Từ 18 đến 718
<b>ELAPSED_TIME</b>	Tổng thời gian bay thực tế	Tổng thời gian thực tế của chuyến bay: AIR_TIME + TAXI_IN + TAXI_OUT (được tính bằng phút)	Từ 14 đến 766

<b>AIR_TIME</b>	Thời gian bay (trên không)	Khoảng thời gian từ lúc bánh rời mặt đất (Wheels Off) đến lúc bánh chạm đất (Wheels On). (tính bằng phút)	Từ 7 đến 690
<b>DISTANCE</b>	Khoảng cách	Khoảng cách giữa hai sân bay (tính bằng dặm - miles)	Từ 21 đến 4983
<b>WHEELS_ON</b>	Thời điểm bánh chạm đất	Thời điểm bánh máy bay chạm mặt đất (hạ cánh)	Từ 1 đến 2400
<b>TAXI_IN</b>	Thời gian lăn bánh vào	Khoảng thời gian từ khi bánh máy bay chạm đất (Wheels On) đến khi đến cổng sân bay đích, được tính bằng phút	Từ 1 đến 248
<b>SCHEDULED_ARRIVAL</b>	Thời điểm đến dự kiến	Thời điểm đến nơi đã lên kế hoạch (được format theo dạng 24 giờ)	Từ 1 đến 2400
<b>ARRIVAL_TIME</b>	Thời điểm đến thực tế	Thời điểm bánh máy bay chạm đất: (WHEELS_ON) + thời gian lăn bánh vào (TAXI_IN) (Thời gian thực tế đến cổng)	Từ 1 đến 2400
<b>ARRIVAL_DELAY</b>	Thời gian trễ đến	Thời gian trễ lúc đến nơi: ARRIVAL_TIME - SCHEDULED_ARRIVAL	Từ -87 đến 1971
<b>DIVERTED</b>	Chuyến bay bị chuyển hướng	Máy bay hạ cánh tại sân bay ngoài lịch trình (1 = Bị chuyển hướng)	0 hoặc 1
<b>CANCELLED</b>	Chuyến bay bị hủy	Chuyến bay bị hủy (1 = Bị hủy)	0 hoặc 1
<b>CANCELLATION_REASON</b>	Lý do hủy chuyến	Lý do hủy chuyến: A - Hàng hàng không; B - Thời tiết; C - Hệ thống Không lưu Quốc gia; D - An ninh	Null/A/B/C/D

<b>AIR_SYSTEM_DELAY</b>	Trễ do hệ thống không lưu	Thời gian trễ do hệ thống không lưu	Từ 0 đến 1134
<b>SECURITY_DELAY</b>	Trễ do an ninh	Thời gian trễ do vấn đề an ninh	Từ 0 đến 573
<b>AIRLINE_DELAY</b>	Trễ do hãng hàng không	Thời gian trễ do hãng hàng không	Từ 0 đến 1971
<b>LATE_AIRCRAFT_DELAY</b>	Trễ do máy bay đến trễ	Thời gian trễ do máy bay thực hiện chuyến bay trước đó đã đến trễ	Từ 0 đến 1331
<b>WEATHER_DELAY</b>	Trễ do thời tiết	Thời gian trễ do thời tiết	Từ 0 đến 1211

## II. Thiết kế NDS



### 1. Thiết kế và đánh giá dạng chuẩn (Normalization Analysis)

Từ cấu trúc bảng của các file nguồn (CSV), ta tiến hành phân tích và đánh giá để đạt dạng chuẩn 3 (3NF) như sau:

#### Bảng AIRLINES (IATA\_CODE, AIRLINE)

Đạt chuẩn BCNF. Mọi thuộc tính đều phụ thuộc trực tiếp vào khóa chính duy nhất (IATA\_CODE).

#### Bảng AIRPORTS (IATA\_CODE, AIRPORT, CITY, STATE, COUNTRY, LATITUDE, LONGITUDE)

Đạt chuẩn BCNF. Các thuộc tính địa lý được xem là thuộc tính đơn trị gắn liền với sân bay. Mọi thuộc tính đều phụ thuộc trực tiếp vào khóa chính duy nhất (IATA\_CODE) (trong ngữ cảnh quản lý địa điểm).

## Bảng FLIGHTS

Không đạt dạng chuẩn 3 (3NF) vì vi phạm các vi phạm ràng buộc:

- Phụ thuộc hàm không khóa: ORIGIN\_AIRPORT, DESTINATION\_AIRPORT  
=> DISTANCE.  
  - ⇒ Tách thành bảng riêng NDS\_Distance.
- Phụ thuộc bắc cầu: CANCELLED (Trạng thái) => CANCELLATION\_REASON (Mã lý do).  
  - ⇒ Tách danh mục lý do thành bảng riêng NDS\_Reason.
- Dư thừa dữ liệu thời gian: Các cột YEAR, MONTH, DAY, DAY\_OF\_WEEK lặp lại liên tục.  
  - ⇒ Tách thành bảng chiều thời gian NDS\_Time.
- Loại bỏ các thuộc tính suy diễn:
  - Departure\_time = wheel\_off - taxi\_out
  - Elapsed\_time = air\_time + taxi\_in + taxi\_out
  - Arrival\_time = wheels\_on + taxi\_in

Sau khi tách các bảng và loại bỏ các thuộc tính trên, bảng FLIGHTS giữ lại các khóa ngoại (Foreign Keys) tham chiếu đến các bảng mới => Đạt dạng chuẩn 3.

## 2. Mô tả NDS

### a. NDS\_Airline

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	IATA_Code (PK)	Là mã định danh duy nhất của hãng hàng không.	Char(2)
2	Airline	Tên đầy đủ của hãng hàng không.	Nvarchar(255)

3	Created_Date	Ngày dòng dữ liệu được tạo	Datetime
4	Updated_Date	Ngày dòng dữ liệu được cập nhật	Datetime

### b. NDS\_Airport

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	IATA_Code (PK)	Là mã định danh duy nhất của sân bay	Char(3)
2	Airport	Tên đầy đủ của sân bay	Varchar(255)
3	City	Thành phố mà sân bay phục vụ	Varchar(255)
4	State	Bang/Vùng mà sân bay tọa lạc	Varchar(50)
5	Country	Quốc gia mà sân bay tọa lạc.	Varchar(50)
6	Latitude	Vĩ độ địa lý của sân bay	Float
7	Longitude	Kinh độ địa lý của sân bay	Float
8	Created_Date	Ngày dòng dữ liệu được tạo	Datetime
9	Updated_Date	Ngày dòng dữ liệu được cập nhật	Datetime

### c. NDS\_Reason

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	ReasonID (PK)	Là mã định danh duy nhất một nguyên nhân.	Char(1)
2	Reason	Nguyên nhân hủy chuyến	Nvarchar(255)
8	Created_Date	Ngày dòng dữ liệu được tạo	Datetime
9	Updated_Date	Ngày dòng dữ liệu được cập nhật	Datetime

### d. NDS\_Distance

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Origin_Airport	Sân bay khởi hành	Char(3)
2	Destination_Airport	Sân bay đến	Char(3)

3	Distance	Khoảng cách giữa hai sân bay	Int
4	Created_Date	Ngày dòng dữ liệu được tạo	Datetime
5	Updated_Date	Ngày dòng dữ liệu được cập nhật	Datetime

#### e. NDS\_Time

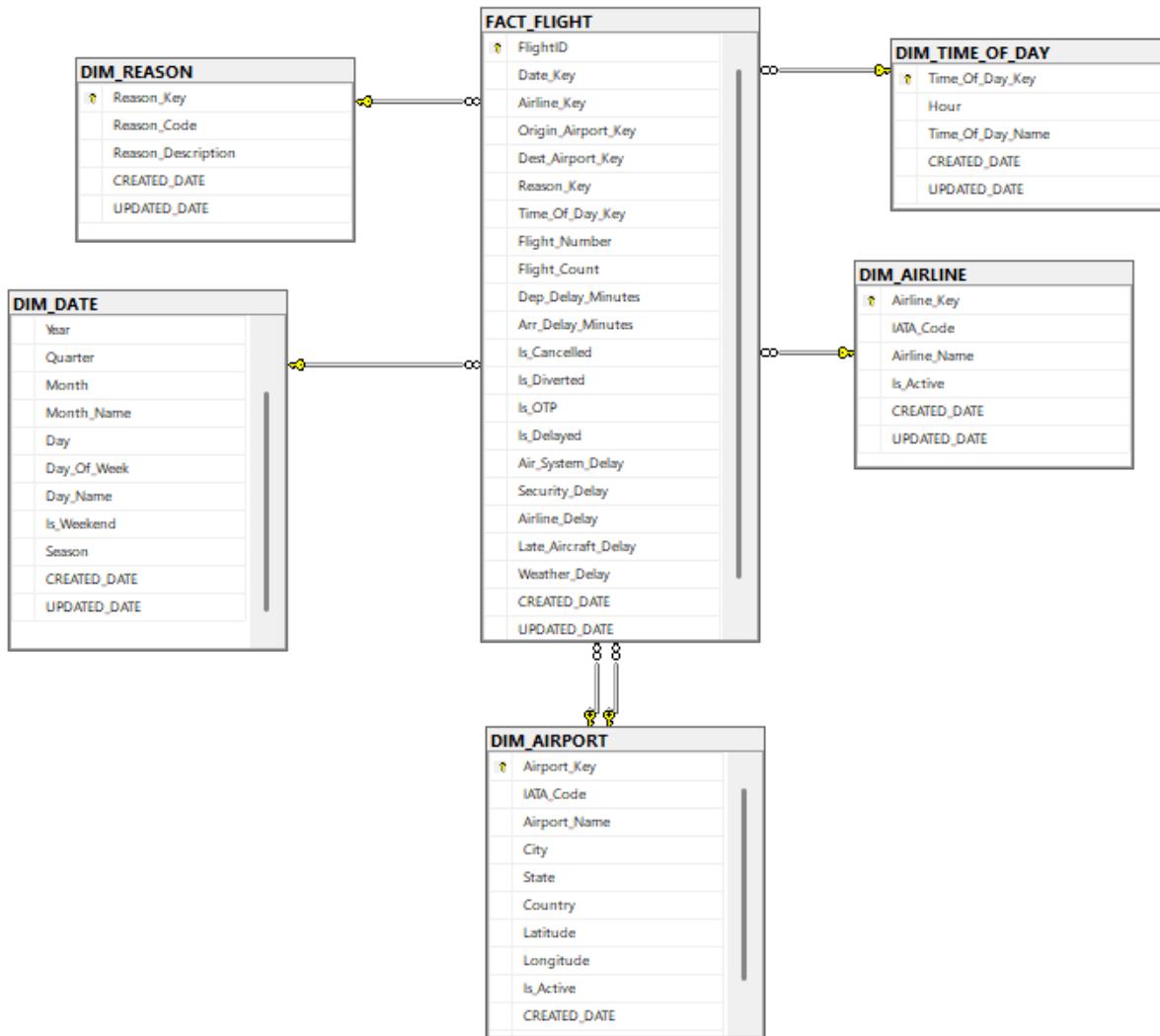
STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	TimeID	ID tự tăng cho các dòng dữ liệu	Int
2	Year	Năm	Int
3	Month	Tháng	Int
4	Day	Ngày	Int
5	Day_Of_Week	Ngày trong tuần	Varchar(50)

#### f. NDS\_Flight

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Flight_ID (PK)	ID tự tăng	Int
2	IATA_Airline (FK)	Mã hãng hàng không	Char(2)
3	Flight_Number	Số hiệu chuyến bay	Int
4	Time_ID (FK)	Thời gian của chuyến bay	Int
5	Tail_Number	Số hiệu máy bay	Varchar(20)
6	Origin_Airport (FK)	Sân bay khởi hành	Char(3)
7	Destination_Airport (FK)	Sân bay đến	Char(3)
8	Scheduled_Departure	Thời điểm khởi hành dự kiến	Time
9	Departure_Delay	Thời gian trễ khởi hành	Int
10	Taxi_Out	Khoảng thời gian lăn bánh ra	Int
11	Wheels_Off	Thời điểm bánh rời mặt đất	Time
12	Scheduled_Time	Thời gian bay dự kiến	Int
13	Air_Time	Thời gian bay (trên không)	Int
14	Wheels_On	Thời điểm bánh chạm đất	Time
15	Taxi_In	Thời gian lăn bánh vào	Int

16	Scheduled_Arrival	Thời điểm đến dự kiến	Time
17	Arrival_Delay	Thời gian trễ đến	Int
18	Diverted	Chuyến bay bị chuyển hướng	Bit
19	Cancelled	Chuyến bay bị hủy	Bit
20	Air_System_Delay	Thời gian trễ do hệ thống không lưu	Int
21	Security_Delay	Thời gian trễ do an ninh	Int
22	Airline_Delay	Thời gian trễ do hàng hàng không	Int
23	Late_Aircraft_Delay	Thời gian trễ do máy bay đến trễ	Int
24	Weather_Delay	Thời gian trễ do thời tiết	Int
25	Cancellation_Reason	Lý do hủy chuyến (FK)	Char(1)
26	Created_Date	Ngày dòng dữ liệu được tạo	Datetime
27	Updated_Date	Ngày dòng dữ liệu được cập nhật	Datetime

### III. Thiết kế DDS



# 1. Thiết kế

## BUỚC 1: Phân tích nghiệp vụ và xác định sự kiện

Dựa vào yêu cầu truy vấn và yêu cầu về dashboard (Tổng số chuyến, OTP, Hủy chuyến, Delay theo nguyên nhân...):

### a. Sự kiện (Event):

- Một chuyến bay đã hoàn tất (hoặc bị hủy) tại một thời điểm cụ thể.

### b. Bối cảnh sự kiện (Dimensions):

- **Ai (Who):** hãng hàng không, máy bay.
- **Ở đâu (Where):** sân bay đi (Origin), sân bay đến (Dest).
- **Khi nào (When):** ngày bay, giờ bay.
- **Tại sao (Why):** Lý do hủy hoặc delay.

### c. Đo lường (Measures):

- **Các dữ liệu có sẵn:** Khoảng cách (Distance), Thời gian trễ (Dep\_Delay\_Minutes, Arr\_Delay\_Minutes), Thời gian bay (Air\_Time), tình trạng hủy (Is\_Cancelled), tình trạng chuyển hướng (Is\_Diverted)
- **Dữ liệu cần phải tính toán**
  - o OTP (Đúng giờ: Arr\_Delay <= 15 phút).

## BUỚC 2: Mô hình hóa (Modeling)

### 1. Thiết kế Chiều (Dimension Design) và SCD

- **Dim\_Date:** phân tích Tháng/Quý/Năm/Mùa.
- **Dim\_Time\_Of\_Date:** phân tích giờ trong ngày.
- **Dim\_Airline**
  - o **Airline\_Name:** Tên hãng hàng không có thể thay đổi do đổi thương hiệu, tái cấu trúc hoặc sáp nhập. Cần lưu lịch sử để đảm bảo thông tin hãng tại thời điểm chuyến bay phát sinh. Áp dụng SCD Type 2.
- **Dim\_Airport:**
  - o **Airport\_Name, City, State:** Các thuộc tính này có thể thay đổi theo thời gian (đổi tên sân bay, điều chỉnh địa giới hành chính). Việc lưu lịch sử giúp đảm bảo dữ liệu phân tích phản ánh đúng bối cảnh tại thời điểm phát sinh chuyến bay. Áp dụng SCD Type 2.
- **Dim\_Reason:**

- **Reason Description:** Danh mục lý do chuyến bay (Weather, Carrier, Security, NAS, Late Aircraft) mang tính cố định, ít thay đổi và không yêu cầu theo dõi lịch sử. Nếu có điều chỉnh chỉ mang tính sửa lỗi hoặc chuẩn hóa tên gọi, không ảnh hưởng đến phân tích dữ liệu trong quá khứ. Vì vậy, áp dụng SCD Type 1 bằng cách cập nhật trực tiếp giá trị khi cần thiết.

## 2. Thiết kế Fact (Fact Design)

- **Độ mịn (Grain):** 1 dòng trong Fact = 1 chuyến bay cụ thể (Flight).
- **Chiều thoái hóa (Degenerate Dimension):**
  - Flight\_Number: Số hiệu chuyến bay.

## 2. Mô tả DDS

### a. Dim\_Date

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Date_Key (PK)	Khóa ngày dạng YYYYMMDD	INT
2	Full_Date	Ngày đầy đủ	DATE
3	Year	Năm	INT
4	Quarter	Quý trong năm	INT
5	Month	Tháng	INT
6	Month_Name	Tên tháng	VARCHAR(20)
7	Day	Ngày trong tháng	INT
8	Day_Of_Week	Thứ trong tuần (dạng số)	INT
9	Day_Name	Tên thứ trong tuần	VARCHAR(20)
10	Is_Weekend	Xác định ngày cuối tuần	BIT
11	Season	Mùa trong năm	VARCHAR(20)
12	Created_Date	Ngày tạo dòng dữ liệu	DATETIME
13	Updated_Date	Ngày cập nhật dòng dữ liệu	DATETIME

### b. Dim\_Time\_Of\_Date

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Time_Of_Day_Key (PK)	Khóa thay thế định danh khung giờ	INT
2	Hour	Giờ trong ngày (0–23)	INT
3	Time_Of_Day_Name	Tên khung giờ (Night, Morning, ...)	VARCHAR(50)
4	Created_Date	Ngày tạo dòng dữ liệu	DATETIME
5	Updated_Date	Ngày cập nhật dòng dữ liệu	DATETIME

### c. Dim\_Airline

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Airline_Key (PK)	Khóa thay thế định danh hãng hàng không trong DDS	INT
2	IATA_Code	Mã IATA định danh hãng hàng không (business key từ NDS)	CHAR(2)
3	Airline_Name	Tên đầy đủ của hãng hàng không	NVARCHAR(255)
4	Is_Active	Trạng thái hoạt động của hãng hàng không	BIT
5	Created_Date	Ngày dòng dữ liệu được tạo trong DDS	DATETIME
6	Updated_Date	Ngày dòng dữ liệu được cập nhật trong DDS	DATETIME

### d. Dim\_Airport

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	Airport_Key (PK)	Khóa thay thế định danh sân bay trong DDS	INT

<b>2</b>	IATA_Code	Mã IATA định danh duy nhất sân bay (business key từ NDS)	CHAR(3)
<b>3</b>	Airport_Name	Tên đầy đủ của sân bay	VARCHAR(255)
<b>4</b>	City	Thành phố mà sân bay phục vụ	VARCHAR(255)
<b>5</b>	State	Bang/Vùng nơi sân bay tọa lạc	VARCHAR(50)
<b>6</b>	Country	Quốc gia nơi sân bay tọa lạc	VARCHAR(50)
<b>7</b>	Latitude	Vĩ độ địa lý của sân bay	FLOAT
<b>8</b>	Longitude	Kinh độ địa lý của sân bay	FLOAT
<b>9</b>	Is_Active	Trạng thái hoạt động của sân bay	BIT
<b>10</b>	Created_Date	Ngày dòng dữ liệu được tạo trong DDS	DATETIME
<b>11</b>	Updated_Date	Ngày dòng dữ liệu được cập nhật trong DDS	DATETIME

### e. Dim\_Reason

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
<b>1</b>	Reason_Key (PK)	Khóa thay thế định danh nguyên nhân trong DDS	INT
<b>2</b>	Reason_Code	Mã nguyên nhân huỷ/trễ chuyến bay	CHAR(1)
<b>3</b>	Reason_Description	Mô tả chi tiết nguyên nhân huỷ chuyến	NVARCHAR(255)
<b>4</b>	Created_Date	Ngày dòng dữ liệu được tạo trong DDS	DATETIME
<b>5</b>	Updated_Date	Ngày dòng dữ liệu được cập nhật trong DDS	DATETIME

## f. Fact\_Flight

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	FlightID (PK)	Khóa chính của bảng	INT
2	Date_Key (FK)	Tham chiếu ngày chuyến bay	INT
3	Airline_Key (FK)	Tham chiếu hãng hàng không	INT
4	Origin_Airport_Key (FK)	Tham chiếu sân bay khởi hành	INT
5	Dest_Airport_Key (FK)	Tham chiếu sân bay đến	INT
6	Reason_Key (FK)	Tham chiếu nguyên nhân huỷ	INT
7	Time_Of_Day_Key (FK)	Tham chiếu khung giờ bay	INT
8	Flight_Number	Số hiệu chuyến bay (Degenerate Dimension)	INT
9	Flight_Count	Số chuyến bay (luôn = 1) (dùng để đếm số chuyến bay)	INT
10	Dep_Delay_Minutes	Thời gian trễ khởi hành	INT
11	Arr_Delay_Minutes	Thời gian trễ đến nơi	INT
12	Is_Cancelled	Xác định chuyến bay bị huỷ	INT
13	Is_Diverted	Xác định chuyến bay chuyển hướng	INT
14	Is OTP	Đúng giờ ( $\text{Arr\_Delay} \leq 15$ phút)	INT
15	Air_System_Delay	Trễ do hệ thống không lưu	INT
16	Security_Delay	Trễ do an ninh	INT
17	Airline_Delay	Trễ do hãng hàng không	INT
18	Late_Aircraft_Delay	Trễ do máy bay đến muộn	INT
19	Weather_Delay	Trễ do thời tiết	INT

20	Created_Date	Ngày tạo dòng dữ liệu	DATETIME
21	Updated_Date	Ngày cập nhật dòng dữ liệu	DATETIME

## IV. Thiết kế Metadata

### 1. ds\_table\_type

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	table_type_key (PK)	Khóa định danh (ID) duy nhất loại bảng	Int
2	table_type	Tên loại bảng	Varchar(20)
3	description	Mô tả	Nvarchar(255)
4	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
5	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

### 2. ds\_table

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	table_key (PK)	Khóa định danh (ID) duy nhất một bảng	Int
2	name	Tên vật lý của bảng trong cơ sở dữ liệu	Varchar(255)
3	entity_type (FK)	Mã chỉ định loại bảng. Là khóa ngoại tham chiếu đến bảng ds_table_type.	Int
4	data_store	Thuộc loại data store	Varchar(20)
5	description	Mô tả chi tiết về mục đích, nội dung và nguồn gốc của bảng	Nvarchar(255)
6	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
7	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

### 3. ds\_column

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	column_key (PK)	Khóa định danh (ID) duy nhất một cột	Int
2	table_key (FK)	Khóa ngoại (ID) của bảng mà cột này thuộc về. Dùng để liên kết với bảng ds_table	Int
3	column_name	Tên vật lý của cột	Varchar(255)
4	data_type	Kiểu dữ liệu của cột	Varchar(255)
5	is_PK	Cờ chỉ báo (Flag) cho biết cột này có phải là Khóa chính hay không	Bit
6	is_FK	Cờ chỉ báo cho biết cột này có phải khóa ngoại hay không	Bit
7	is_null	Cờ chỉ báo (Flag) cho biết cột này có cho phép giá trị NULL hay không	Bit
8	is_identity	Cờ chỉ báo (Flag) cho biết cột này có được thiết lập là cột Tự động tăng.	Bit
9	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
10	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

### 4. status\_table

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	status_key (PK)	Khóa định danh (ID) duy nhất cho một trạng thái	Int
2	status	Tên ngắn, dễ đọc của trạng thái	Varchar(50)
10	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
11	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

## 5. data\_flow

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	data_flow_key	Khóa định danh (ID) duy nhất cho luồng dữ liệu (Data Flow) này	Int
2	name	Tên ngắn, dễ đọc của luồng dữ liệu/tác vụ ETL	Varchar(50)
3	description	Mô tả chi tiết về mục đích và các bước nghiệp vụ mà luồng dữ liệu này thực hiện	Nvarchar(255)
4	source	Nguồn dữ liệu mà luồng này trích xuất (Extract) hoặc đọc vào.	Varchar(50)
5	target	Đích dữ liệu mà luồng này ghi (Load) vào. Thường là tên bảng đích sau quá trình chuyển đổi	Varchar(50)
6	transformation	Mô tả chi tiết các quy tắc chuyển đổi (Transformation rules) được áp dụng cho dữ liệu trong luồng này	Nvarchar(255)
7	Status (FK)	Trạng thái hiện tại của luồng dữ liệu/tác vụ này, tham chiếu đến Status Table.	Int
8	LSET	Last Successful Execution Time (Thời điểm chạy thành công gần nhất).	Datetime
9	CET	Current Execution Time (Thời điểm chạy hiện tại/gần nhất). Lưu lại thời gian bắt đầu hoặc kết thúc lần chạy gần nhất	Datetime
10	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
11	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

## 6. dq\_rule\_category

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	category_id (PK)	Khóa chính, mã định danh duy nhất cho danh mục kiểm tra (ví dụ: I, C, D).	Char(1)
2	description	Mô tả chi tiết về giai đoạn áp dụng quy tắc (ví dụ: Kiểm tra dữ liệu đầu vào tại Stage).	Varchar(100)

## 7. dq\_rule\_action

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	action_id (PK)	Khóa chính, mã định danh cho hành động xử lý (ví dụ: R, A, F).	Char(1)
2	description	Mô tả về chiến lược xử lý lỗi (ví dụ: Từ chối bản ghi, Cho phép nạp kèm cảnh báo, hoặc Tự động sửa).	Varchar(50)

## 8. dw\_user

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	user_key	Khóa định danh duy nhất cho người dùng hệ thống	Int
2	name	Họ và tên đầy đủ của user	Nvarchar(100)
3	department	Phòng ban hoặc bộ phận công tác	Nvarchar(100)
4	role	Vai trò trong dự án	Nvarchar(100)
5	email_address	Địa chỉ email để gửi thông báo lỗi DQ	Varchar(255)
6	user_group_key	Dùng để phân nhóm người dùng	Int

## 9. dq\_rule

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
-----	------------	---------	--------------

1	rule_key (PK)	Khóa định danh duy nhất cho quy tắc chất lượng dữ liệu	Int
2	rule_name	Tên ngắn gọn của quy tắc (ví dụ: VAL_IATA_CODE)	Nvarchar(255)
3	description	Mô tả chi tiết logic kiểm tra và điều kiện vi phạm	Nvarchar(Max)
4	rule_type	Phân loại mức độ vi phạm: Error hoặc Warning	Char(1)
5	rule_category (FK)	Danh mục kiểm tra: I (Stage), C (NDS), D (DDS)	Char(1)
6	risk_level	Mức độ rủi ro đối với nghiệp vụ khi quy tắc bị vi phạm (thang điểm 1-5)	Int
7	status	Trạng thái hoạt động của quy tắc	Varchar(7)
8	Action (FK)	Hành động hệ thống sẽ thực hiện khi vi phạm: R (Reject), A (Allow), F (Fix)	Char(1)
	table_key (FK)	FK liên kết đến bảng chịu tác động trong hệ thống metadata (ds_table)	Int
10	create_timestamp	Ngày dòng dữ liệu được tạo	Datetime
11	update_timestamp	Ngày dòng dữ liệu được cập nhật	Datetime

## 10.dq\_notification

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	notification_key (PK)	Khóa chính định danh duy nhất cho mỗi thiết lập thông báo.	Int
2	rule_key	Khóa ngoại liên kết với quy tắc cụ thể trong bảng dq_rule	Int
3	recipient_type	Phân loại đối tượng nhận thông báo: I (Individual - Cá nhân) hoặc G (Group - Nhóm).	Char(1)

4	recipient_id	ID của người dùng (từ bảng dw_user) hoặc ID nhóm người dùng tương ứng với loại đối tượng.	Int
5	method	Phương thức gửi thông báo: E (Email) hoặc S (SMS/Text message).	Char(1)
6	last_notified	Ghi lại mốc thời gian cuối cùng mà hệ thống gửi thông báo cho quy tắc này.	Datetime

## 11.usage\_log

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	usage_key (PK)	Khóa định danh duy nhất cho mỗi lần truy cập dữ liệu	Int
2	user_key (FK)	FK liên kết đến người dùng thực hiện truy cập (dw_user)	Int
3	object_key (FK)	FK liên kết đến đối tượng được truy cập	Int
4	access_via	Giao diện / công cụ dùng để truy vấn	Nvarchar(50)
5	timestamp	Thời điểm chính xác sự kiện truy cập diễn ra	Datetime

## V. Business rules

### 1. NDS\_Reason

Rule	Mô tả
BR-R1	ReasonID ∈ {A,B,C,D}.
BR-R2	Nếu CANCELLED = 0 thì ReasonID phải NULL.

### 2. NDS\_Distance

Rule	Mô tả
BR-D1	Cặp (ORIGIN, DEST) là unique.
BR-D2	Distance > 0, không âm.
BR-D3	ORIGIN ≠ DEST.

### 3. NDS\_Time

Rule	Mô tả
BR-T1	Bộ (YEAR, MONTH, DAY) unique → sinh TimeID.
BR-T2	DAY_OF_WEEK chỉ gồm các giá trị (Monday - Sunday)

### 4. NDS\_Flight

Rule	Mô tả
BR-N1	Composite PK = (IATA_Airline, Flight_Number, TimeID, Origin_Airport).
BR-N2	Foreign key phải khớp 100% (Airport/Airline/Time/Reason).
BR-N3	Nếu DIVERTED = 1: Destination_Airport vẫn phải có trong domain Airport.
BR-N4	Delay < 0 hợp lệ (early arrival). Không được convert thành 0.
BR-N5	Created_Date auto set lúc load.
BR-N6	Updated_Date set khi record bị update theo LSET/CET.

### 5. Dimensions

Rule	Mô tả
BR-DD1	SCD Type 2 Integrity: Với Dim_Airline/Dim_Airport, tại một thời điểm chỉ tồn tại duy nhất một dòng có Is_Active = 1 cho mỗi Business Key.
BR-DD2	SCD Type 1 Strategy: Với Dim_Reason, khi có thay đổi mô tả, thực hiện ghi đè (Overwrite) trực tiếp lên dữ liệu cũ, không tạo dòng mới.

<b>BR-DD3</b>	Surrogate Key: Dim_Airline, Dim_Airport, Dim_Reason phải sử dụng khóa thay thế (Surrogate Key) dạng số nguyên tự tăng làm Primary Key, không dùng Business Key.
<b>BR-DD4</b>	Date Dimension: Bảng Dim_Date phải bao gồm đầy đủ các ngày trong khoảng thời gian dữ liệu, không được phép thiếu ngày .

## 6. Fact\_Flight

Rule	Mô tả
<b>BR-DF1</b>	Referential Integrity: Mọi khóa ngoại (Airline_Key, Airport_Key,...) trong Fact phải Lookup thành công (Match) với bảng Dim tương ứng.
<b>BR-DF2</b>	Null Measure Handling: Các giá trị đo lường (Delay, Time) nếu là NULL hoặc rỗng ở nguồn phải convert thành 0 để tính toán.
<b>BR-DF3</b>	OTP Logic: Is OTP = 1 khi (Is_Cancelled = 0 VÀ Arr_Delay < 15). Nếu Is_Cancelled = 1 thì Is OTP là 0.
<b>BR-DF4</b>	Boolean Normalization: Các cờ trạng thái (Cancelled, Diverted) phải lưu dưới dạng số nguyên (1/0) thay vì True/False để hỗ trợ hàm SUM.

# VI. Mapping giữa các tầng dữ liệu

Source → Stage → NDS

## 1. Bảng mapping

### 1.1. Mapping Airline

Source (airlines.csv)	Stage (STG_Airline)	NDS (NDS_Airline)	Transformation	Giải thích
IATA_CODE	Iata_Code	Iata_Code (PK)	TRIM → UPPER	Chuẩn hóa về 2 ký tự viết hoa
AIRLINE	Airline_Name	Airline_Name	Chuyển đổi kiểu dữ liệu	Loại bỏ ký tự thừa, chuẩn hóa tên hãng
–	–	Created_Date	GETDATE()	Timestamp khi insert
–	–	Updated_Date	GETDATE()	Timestamp khi update

## 1.2. Mapping Airport

Source (airports.csv)	Stage (STG_Airport)	NDS (NDS_Airport)	Transformation	Giải thích
IATA_CODE	Iata_Code	Iata_Code (PK)	TRIM → UPPER	Chuẩn hóa 3 ký tự viết hoa
AIRPORT	Airport_Name	Airport_Name	Chuyển đổi kiểu dữ liệu	Chuẩn hóa tên sân bay
CITY	City_Name	City_Name	Chuyển đổi kiểu dữ liệu	Chuẩn hóa tên thành phố
STATE	State_Code	State_Code	Chuyển đổi kiểu dữ liệu	Giữ nguyên mã bang
COUNTRY	Country_Name	Country_Name	Chuyển đổi kiểu dữ liệu	Chuẩn hóa quốc gia
LATITUDE	Latitude	Latitude	CAST AS FLOAT	Kiểm tra miền hợp lệ
LONGITUDE	Longitude	Longitude	CAST AS FLOAT	Kiểm tra miền hợp lệ
–	–	Created_Date	GETDATE()	Timestamp insert
–	–	Updated_Date	GETDATE()	Timestamp update

## 1.3. Mapping Time

Source (flights_batch.cs v)	Stage (STG_Flight)	NDS (NDS_Tim e)	Transformation	Giải thích
YEAR	Year_Value	Year	CAST AS INT	
MONTH	Month_Value	Month	CAST AS INT	
DAY	Day_Value	Day	CAST AS INT	
DAY_OF_WEEK	Day_Of_Wee k_Num	Day_Of_W eek	Derived Column: 1 → "Monday",	Chuyển số → tên ngày tiếng Anh

			2 → "Tuesday", ... 7 → "Sunday"	
(YEAR,MONTH, DAY)	—	Time_Id (PK)	Lookup NDS_Time trên 3 trường trên → Match: trả về Time_Id hiện có → No Match: INSERT + sinh Time_Id (IDENTITY)	
—	—	Created_Date	GETDATE()	Timestamp insert
—	—	Updated_Date	GETDATE()	Timestamp update

#### 1.4. Mapping Reason

Source (flights_batch.csv)	Stage (STG_Flight)	NDS (NDS_Reason)	Transformation	Giải thích
CANCELLATION_REASON	Cancellation_Reason	Reason_Id	TRIM, UPPER	Chuẩn hóa mã lý do
—	—	Reason_Name	MAP STATIC	Gán text mô tả (A,B,C, D)
—	—	Created_Date	GETDATE()	Timestamp insert
—	—	Updated_Date	GETDATE()	Timestamp update

#### 1.5. Mapping AirportCodeMapping

Source (Airportmapping)	Stage (STG_Flight)	Transformation
DOT_ID	DOT_Code	TRIM
IATA	IATA_Code	TRIM → UPPER

## 1.6. Mapping Flight

Source (flights_batch.csv)	Stage (STG_Flight)	NDS (NDS_Flight)	Transformation	Giải thích
AIRLINE	Iata_Airline	Iata_Airline (PK, FK)	TRIM → UPPER → DT_STR(2)	Chuẩn hóa mã hãng 2 ký tự
FLIGHT_NUMBE R	Flight_Number	Flight_Number (PK)	CAST AS INT	Chuẩn kiểu dữ liệu
YEAR + MONTH + DAY + DAY_OF_WEEK	Year_Value + Month_Value + Day_Value + Day_Of_Wee k	Time_Id (PK, FK)	Lookup NDS_Time bằng 4 cột (Year_Value, Month_Value, Day_Value, Day_Of_Week_T ext) → trả về Time_Id hiện có hoặc sinh mới nếu chưa tồn tại	—
TAIL_NUM	Tail_Number	Tail_Number	TRIM → DT_STR(20) → ISNULL/blank → 'N/A'	Chuẩn hóa
ORIGIN_AIRPORT	Origin_Airpor t	Origin_Airpor t	Conditional Split → nếu là DOT code → LEFT JOIN STG_AirportCod eMapping → lấy	Xử lý cả mã DOT và IATA

			IATA nếu đã IATA → giữ nguyên → UPPER	
DESTINATION_AI RPORT	Destination_Airport	Destination_Airport	Tương tự ORIGIN	—
SCHEDULED_DE PARTURE	Scheduled_Departure	Scheduled_Departure	HHMM → TIME	Chuyển định dạng
DEPARTURE_TIM E	Departure_Time	Departure_Time	HHMM → TIME	—
DEPARTURE_DE LAY	Departure_Delay	Departure_Delay	CAST INT	—
TAXI_OUT	Taxi_Out	Taxi_Out	CAST INT	—
WHEELS_OFF	Wheels_Off	Wheels_Off	HHMM → TIME	—
SCHEDULED_TIM E	Scheduled_Time	Scheduled_Time	CAST INT	—
ELAPSED_TIME	Elapsed_Time	Elapsed_Time	CAST INT	—
AIR_TIME	Air_Time	Air_Time	CAST INT	—
WHEELS_ON	Wheels_On	Wheels_On	HHMM → TIME	—
TAXI_IN	Taxi_In	Taxi_In	CAST INT	—
SCHEDULED_AR RIVAL	Scheduled_Arrival	Scheduled_Arrival	HHMM → TIME	—
ARRIVAL_TIME	Arrival_Time	Arrival_Time	HHMM → TIME	—
ARRIVAL_DELAY	Arrival_Delay	Arrival_Delay	CAST INT	—
CANCELLED	Cancelled	Cancelled	CAST BIT	—
DIVERTED	Diverted	Diverted	CAST BIT	—
CANCELLATION_ REASON	Cancellation_Reason	Cancellation_Reason	TRIM + LOOKUP DIM	Chuẩn hóa & map lý do
AIR_SYSTEM_DE LAY	Air_system_Delay	Air_system_Delay	ISNULL hoặc blank → 0 → CAST AS INT	

SECURITY_DELAY	Security_delay	Security_delay	ISNULL hoặc blank → 0 → CAST AS INT	
AIRLINE_DELAY	Airline_delay	Airline_delay	ISNULL hoặc blank → 0 → CAST AS INT	
LATE_AIRCRAFT_DELAY	Late_Aircraft_delay	Late_Aircraft_delay	ISNULL hoặc blank → 0 → CAST AS INT	
WEATHER_DELAY	Weather_Delay	Weather_Delay	ISNULL hoặc blank → 0 → CAST AS INT	
	—	Created_Date	GETDATE()	Timestamp insert

## 1.7. Mapping Distance

Source (flights_batch. csv)	Stage (STG_Flight)	NDS (NDS_Distanc e)	Transformati on	Giải thích
ORIGIN_AIRPORT	Origin_Airport	Origin_Airport (PK)	Map DOT→IATA → UPPER	
DESTINATIO N_AIRPORT	Destination_ Airport	Destination_A irport (PK)	Map DOT→IATA → UPPER	
DISTANCE	Distance	Distance	CAST AS INT → ISNULL → 0 Aggregate: lấy MAX(Distan ce) cho mỗi cặp (Origin, Dest)	Tránh duplicate pair, lấy khoảng cách lớn nhất (đúng nhất)
—	—	Created_Date	GETDATE()	

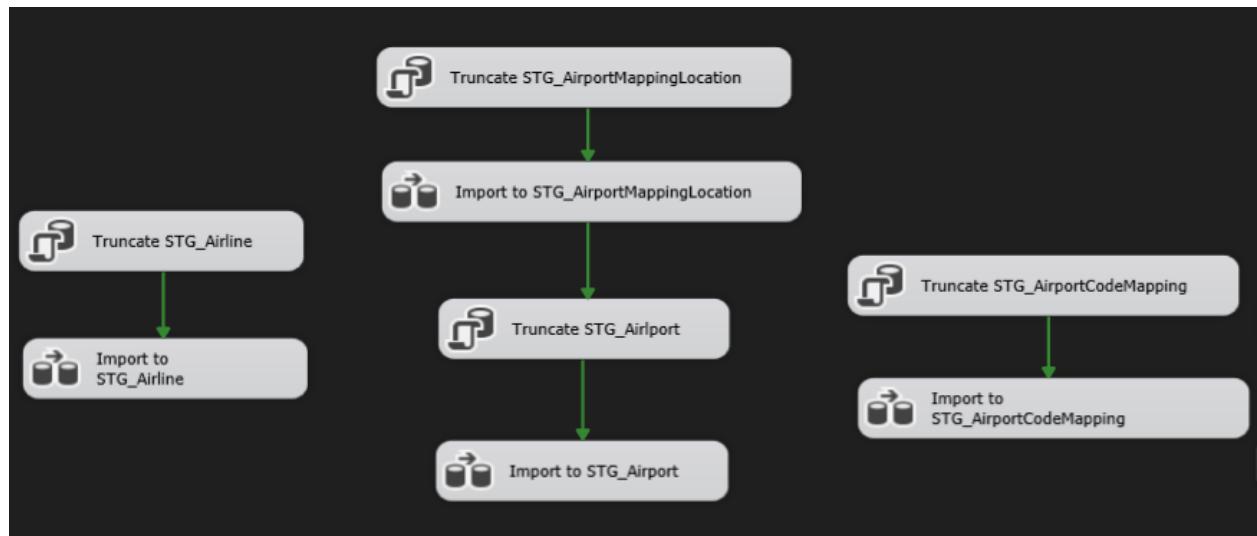
## 2. Phân tích Control và ETL Component

### 2.1. Source -> Stage

#### a. Nhóm Control cho Airline, Airport, AirportCodeMapping, AirportMappingLocation

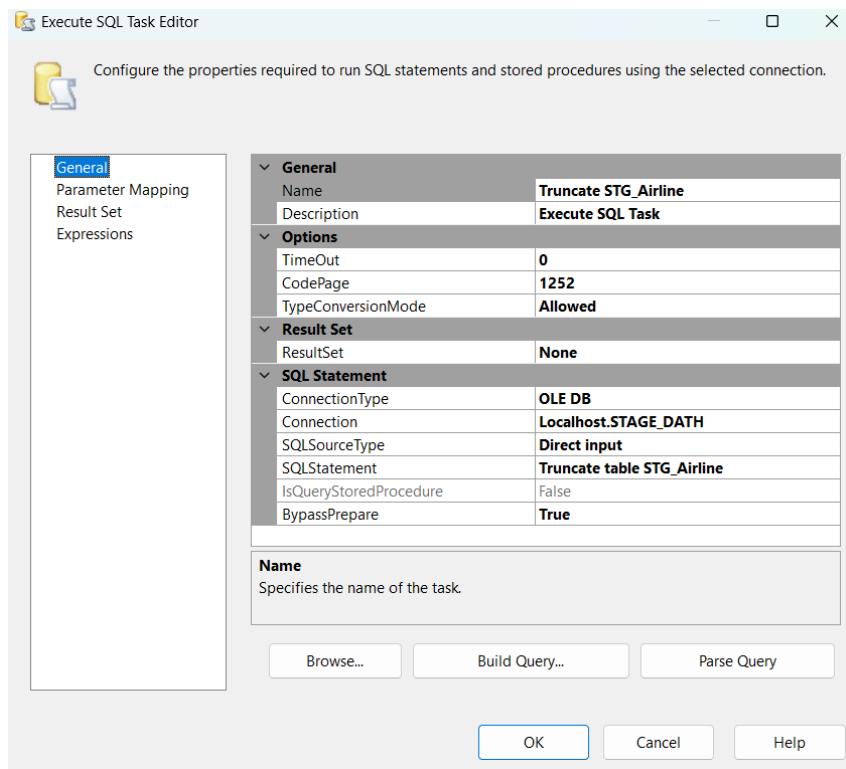
Cấu trúc chung cho mỗi nhóm:

1. Truncate STG\_xxx
2. Import to STG\_xxx



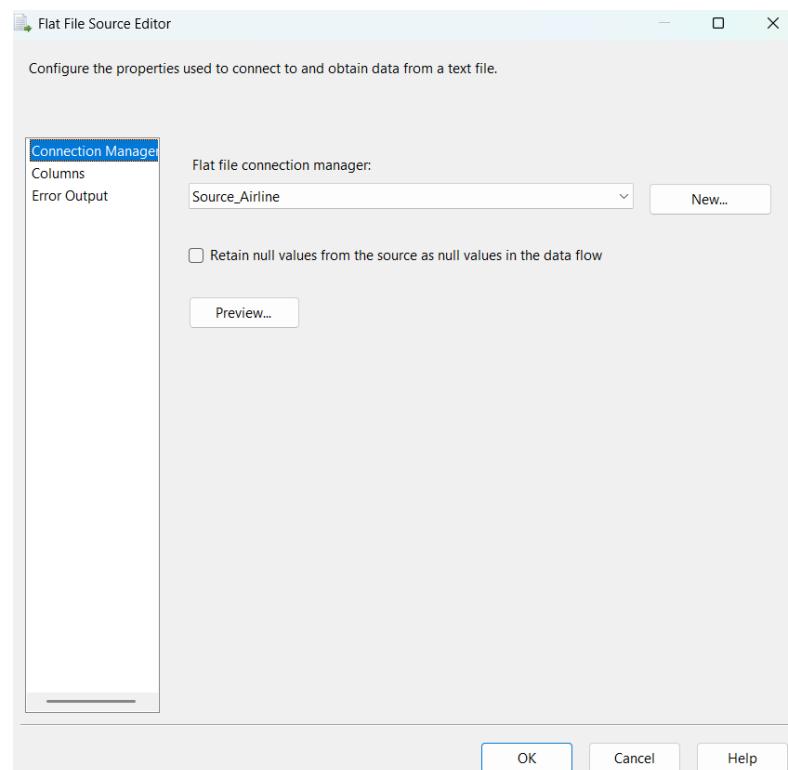
#### Ý nghĩa từng Control:

**Truncate STG\_xxx:** Xóa toàn bộ dữ liệu cũ trong bảng Stage để đảm bảo môi trường sạch trước khi load dữ liệu mới.



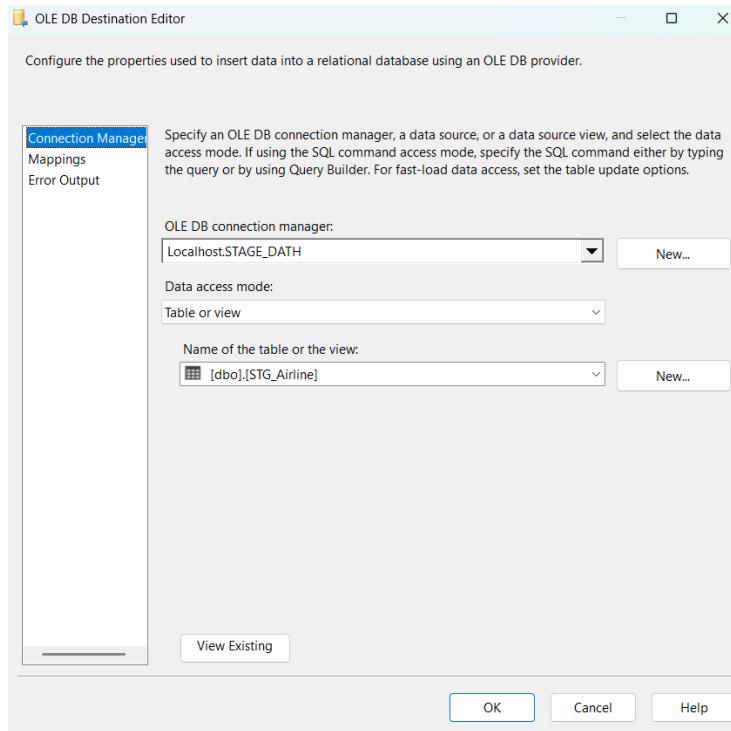
## Import to STG\_xxx

**Bước 1:** Cấu hình Connection Manager trỏ đến file .csv nguồn.



**Bước 2:** Làm sạch và chuẩn hóa các cột (trim, đổi kiểu dữ liệu, validate bắt buộc), mapping đúng schema Stage.

**Bước 3:** Sử dụng Data Flow để chuyển dữ liệu từ file source csv vào bảng Stage.

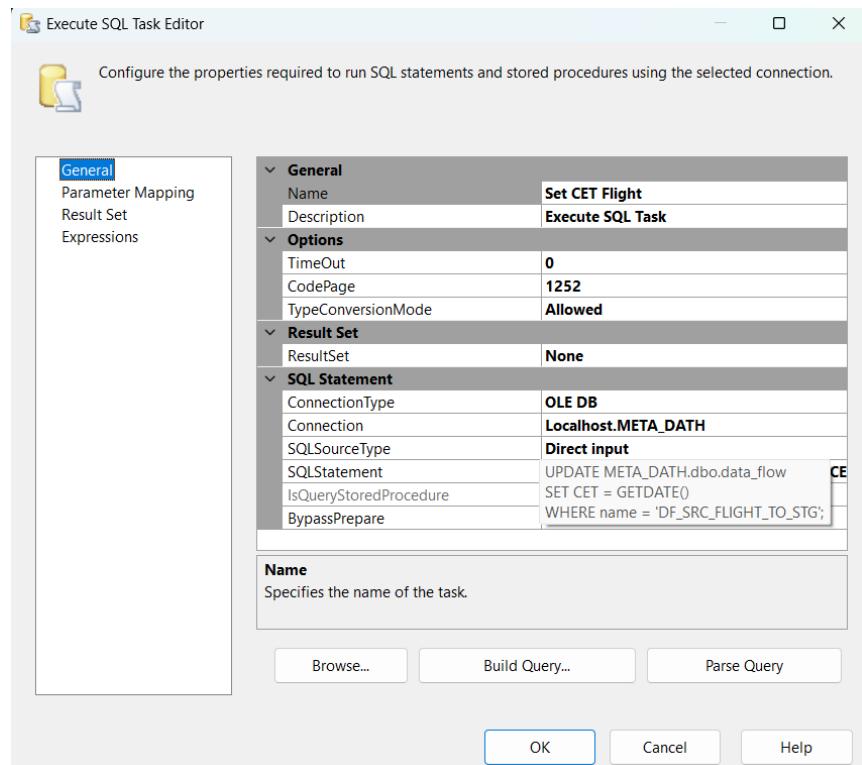


### b. Nhóm Control cho bảng Flight

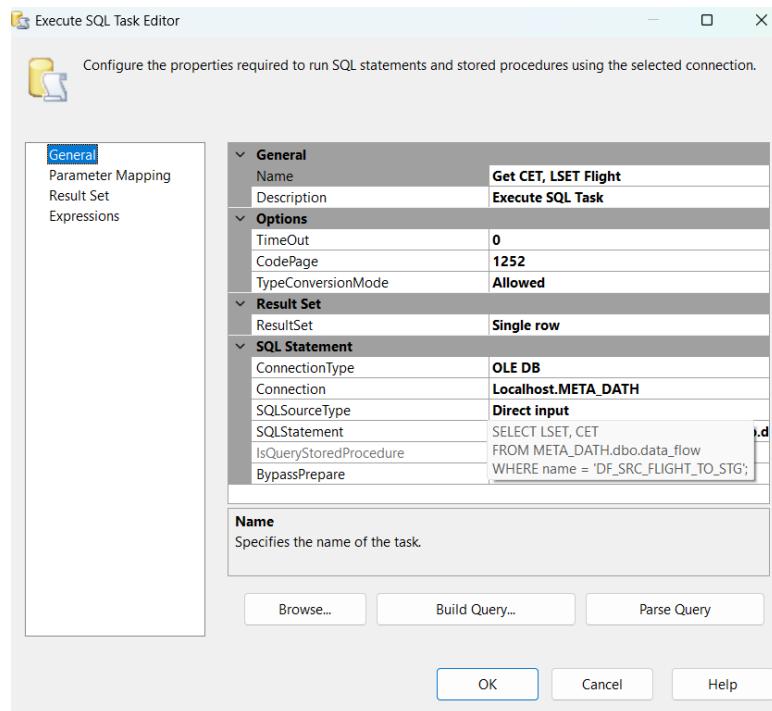


**Ý nghĩa từng Control:** Có 7 bước

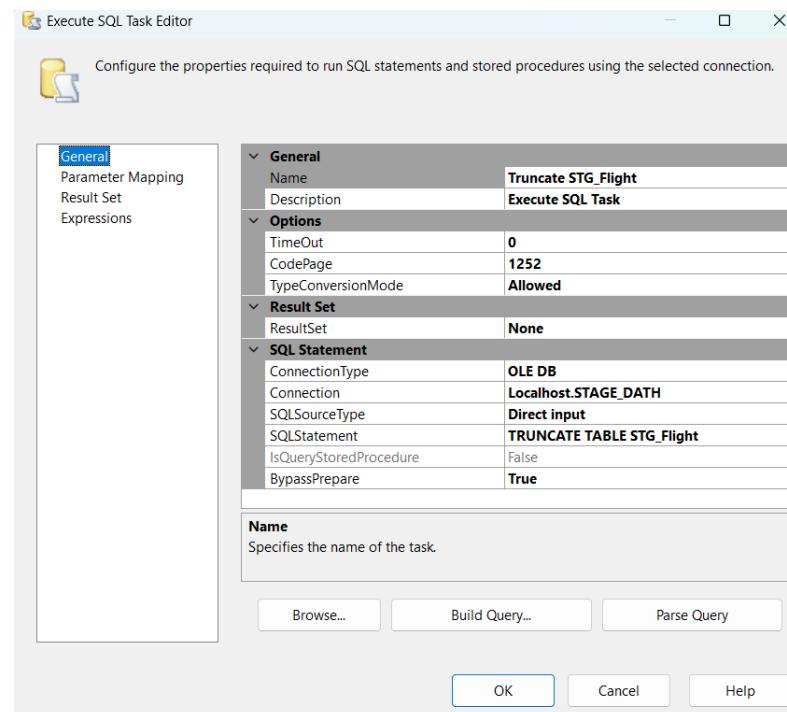
- Set CET Flight:** Cập nhật trường CET (Current Extract Time) của bản ghi quản lý Stage Flight trong bảng metadata META\_DATH.data\_flow bằng timestamp hệ thống hiện tại (GETDATE()).



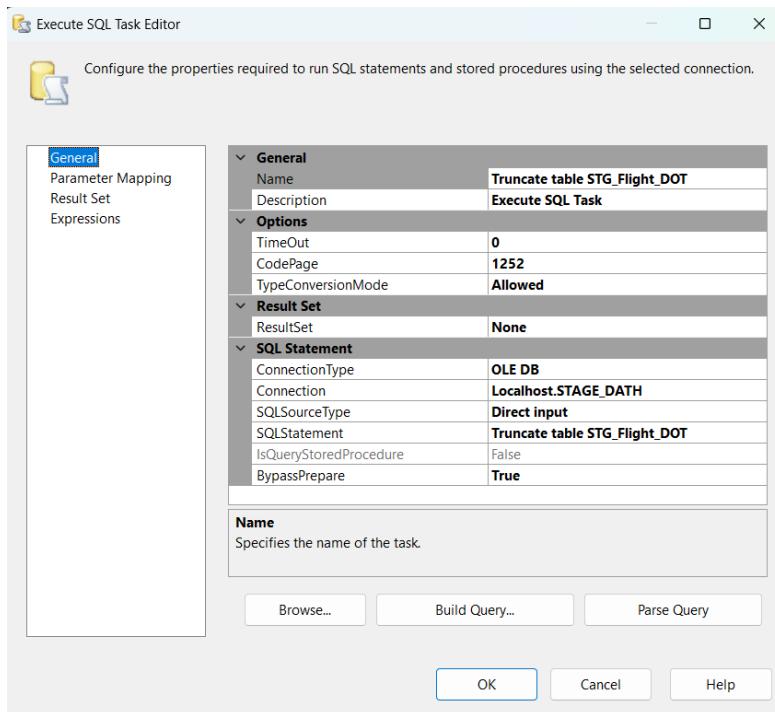
**2. Get CET, LSET Flight:** Truy xuất thời điểm LSET (Last Successful Extract Time) và CET (Current Extract Time) của Stage Flight từ CSDL Metadata. Đây là giá trị lần extract thành công gần nhất và thời điểm khởi tạo package hiện tại, dùng để xác định phạm vi dữ liệu mới cần nạp vào Stage. Thực hiện câu lệnh SELECT LSET, CET trên bảng metadata, cấu hình Result Set ở chế độ *Single row*, và gắn kết quả tương ứng vào biến User::LSET và User::CET.



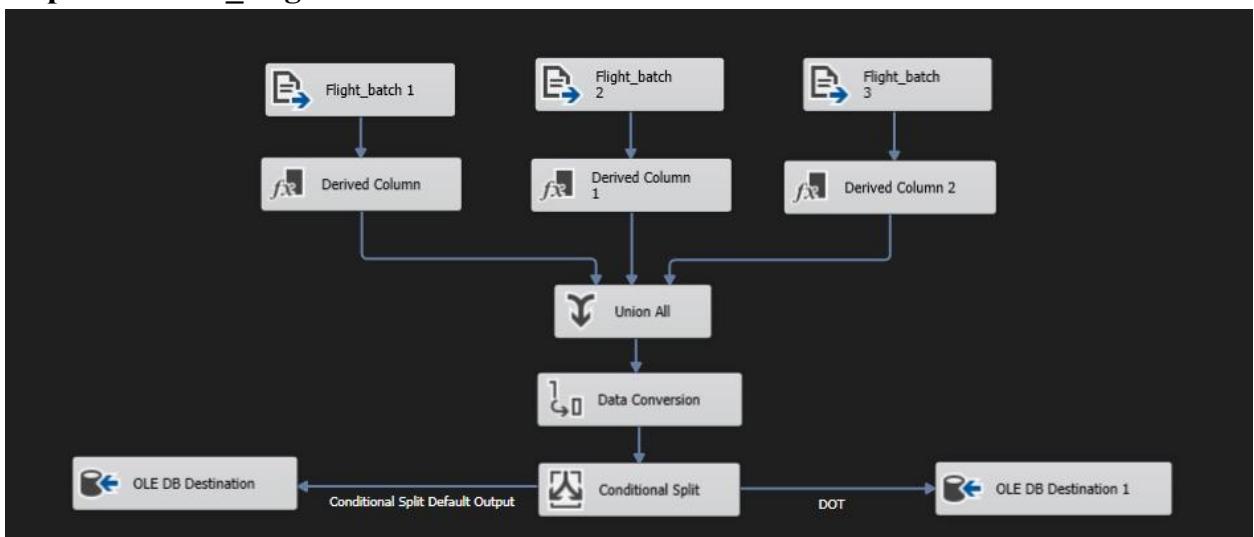
**3. Truncate STG\_Flight:** Làm sạch dữ liệu ở stage trước khi đổ dữ liệu mới vào các bảng.



- 4. Truncate STG\_Flight\_DOT và Truncate STG\_Flight:** làm sạch dữ liệu ở stage trước khi đổ dữ liệu mới vào các bảng.



## 5. Import to STG\_Flight



**Bước 5.1:** Trích xuất dữ liệu nguồn (Source Extraction) từ các file csv  
Flight\_batch 1, Flight\_batch 2, Flight\_batch 3

**Bước 5.2:** Thực hiện các công việc như chuẩn hóa dữ liệu, tạo các cột mới, sửa định dạng, tính toán bằng Derived Column

Thuộc Tính	Chuẩn Hóa
<b>Scheduled_Departure, Departure_Time, Wheels_Off, Scheduled_Arrival, Arrival_Time, Wheels_On</b>	Chuyển thời gian dạng "HHmm" hoặc "HMM" sang định dạng HH:mm:ss chuẩn. Xử lý null/blank thành "00:00:00", và "2400" quy về "00:00:00".
<b>Arrival_Delay, Elapsed_Time, Air_System_Delay, Security_Delay, Airline_Delay, Late_Aircraft_Delay, Weather_Delay, Departure_Delay, Taxi_Out, Taxi_In, Scheduled_Time, Air_Time, Distance</b>	Nếu null hoặc chuỗi trống, gán giá trị "0". Giữ nguyên chuỗi nếu có giá trị (trim whitespace). Dạng số nhưng vẫn giữ dưới dạng string.
<b>Origin_Airport, Destination_Airport</b>	Nếu null hoặc trống, gán chuỗi rỗng "". Nếu có dữ liệu, trim và chuyển thành chữ hoa (UPPER).
<b>Cancellation_Reason</b>	Nếu chuỗi trống, trả về NULL kiểu chuỗi unicode. Nếu có, trim giữ nguyên.
<b>Airline</b>	Chuyển sang kiểu chuỗi fixed length 2 ký tự (DT_STR,2). Không có xử lý null, giữ nguyên.
<b>Tail_Number</b>	Chuyển sang kiểu chuỗi fixed length 20 ký tự (DT_STR,20). Giữ nguyên dữ liệu.
<b>Cancelled, Diverted</b>	Chuyển giá trị "1" thành TRUE, các giá trị khác thành FALSE. Chuẩn hóa kiểu boolean.

**Bước 5.3:** Hợp nhất dữ liệu từ các nguồn sau khi chuẩn hóa bằng Union All

**Bước 5.4:** Dùng Data Conversion để thống nhất tên và thực hiện một số thay đổi kiểu dữ liệu đầu ra phù hợp với các bảng trong stage

Cột Đầu vào (Input Column)	Đầu ra (Output Alias)	Kiểu dữ liệu Đầu ra (Data Type)	Kiểu dữ liệu trong Stage
Cancellation_Reason_Fix_DC	Cancellation_Reason_Fix	string [DT_STR]	string
Late_Aircraft_Delay_Fix_DC	Late_Aircraft_Delay_Fix	four-byte signed integer [DT_I4]	integer
Security_Delay_Fix_DC	Security_Delay_Fix	four-byte signed integer [DT_I4]	integer
Wheels_On_Fix_DC	Wheels_On_Fix	four-byte signed integer [DT_I4]	integer
Weather_Delay_Fix_DC	Weather_Delay_Fix	four-byte signed integer [DT_I4]	integer
Scheduled_Time_Fix_DC	Scheduled_Time_Fix	four-byte signed integer [DT_I4]	integer

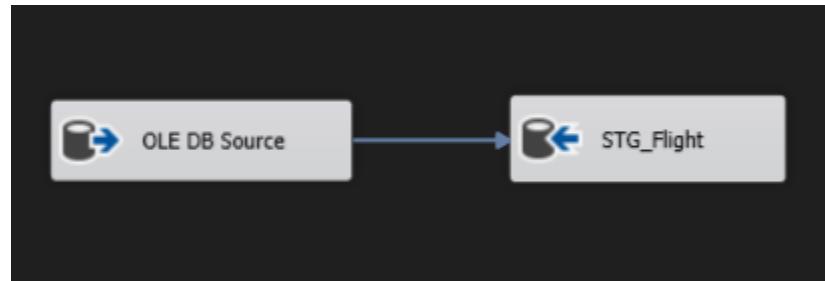
**Bước 5.5:** Dùng Conditional Split để phân loại theo điều kiện cột Destination\_Airport

Điều kiện	Đích đến
Mã sân bay hợp lệ theo IATA	Bảng stage STG_FLIGHT
Mã dạng DOT đặc biệt	Bảng stage STG_FLIGHT_DOT

**Bước 5.6:** Dữ liệu sau phân tách được tải vào 2 bảng stage tương ứng (OLE DB Destination):

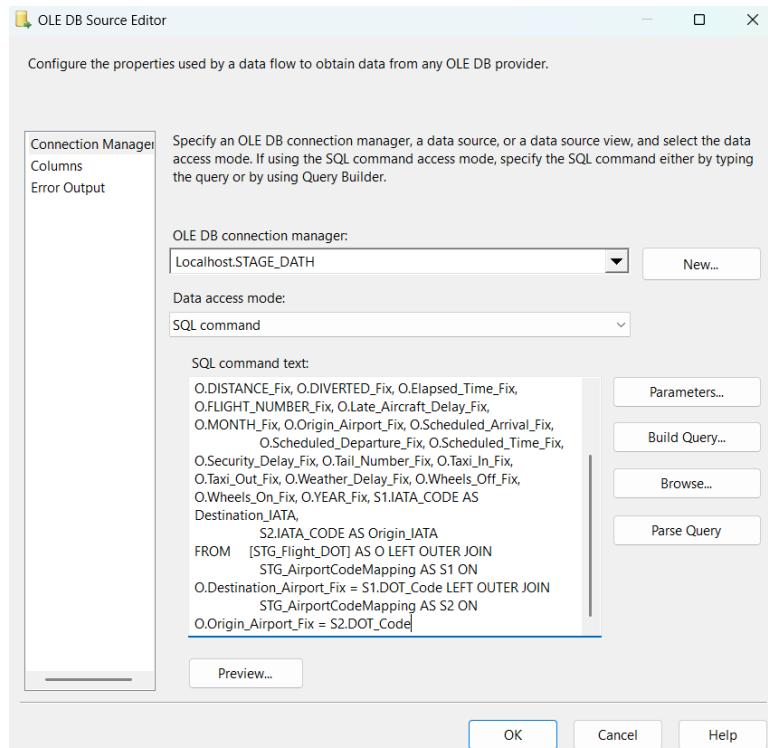
- STG\_FLIGHT cho dữ liệu chuẩn.
- STG\_FLIGHT\_DOT cho dữ liệu có Destination\_Airport dạng DOT.

## 6. Mapping AirportCode



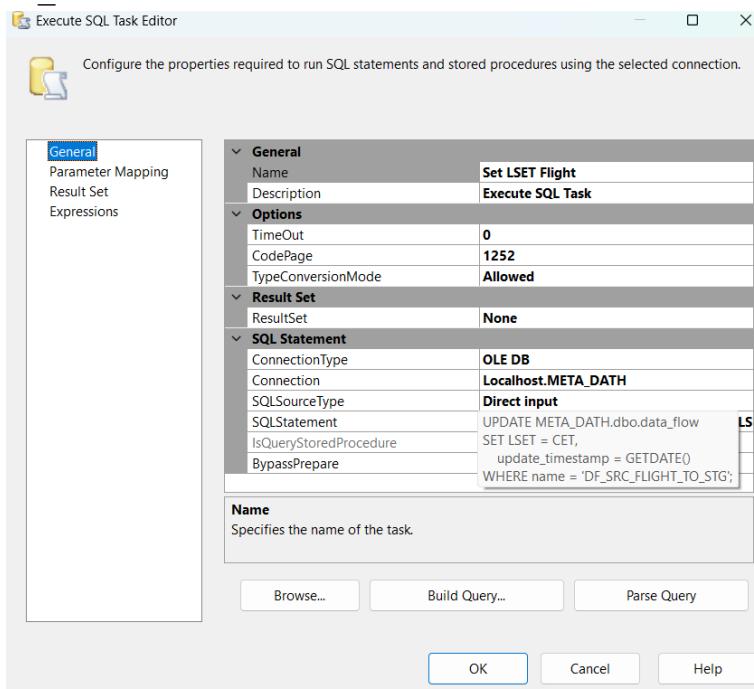
**Bước 6.1:** Lấy mã sân bay chuẩn IATA cho sân bay đến (Destination) và sân bay đi (Origin) và đồ dữ liệu đã chuẩn hóa này vào Stage STG\_Flight, cụ thể:

- Lấy các trường dữ liệu chuyến bay đã được chuẩn hóa trong bảng STG\_Flight\_DOT (biến O).
- Thực hiện LEFT JOIN với bảng STG\_AirportCodeMapping (biến S1) theo điều kiện O.Destination\_Airport\_Fix = S1.DOT\_Code để lấy mã IATA tương ứng của sân bay đến, đặt alias là Destination\_IATA.
- Tương tự, thực hiện LEFT JOIN với bảng STG\_AirportCodeMapping (biến S2) theo điều kiện O.Origin\_Airport\_Fix = S2.DOT\_Code để lấy mã IATA của sân bay đi, alias Origin\_IATA.



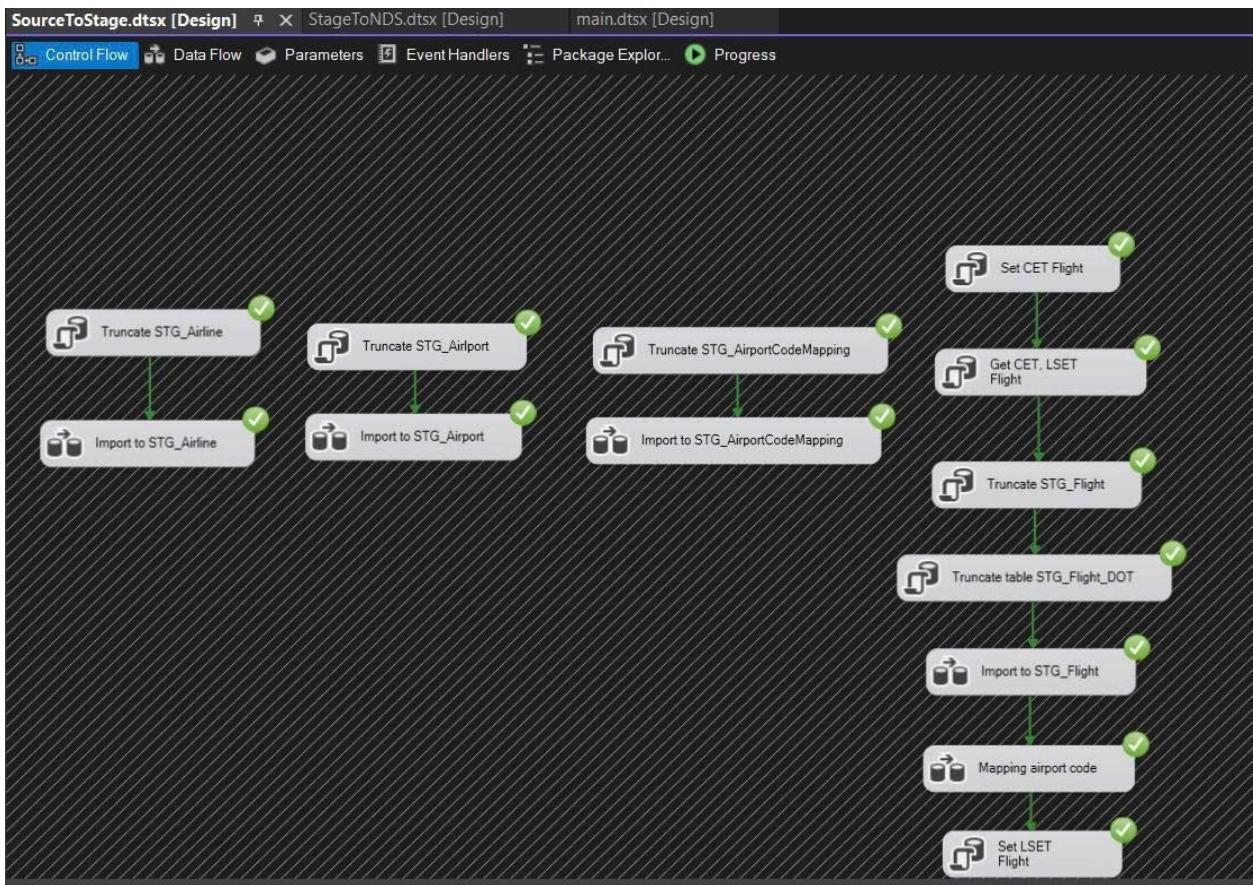
**Bước 6.2:** Đỗ dữ liệu đã lấy mã IATA vào Stage STG\_Flight

7. Set LSET Flight: Cập nhật trường LSET bằng giá trị CET và đồng thời cập nhật thời gian chạy package (update\_timestamp) bằng GETDATE() trong bảng quản lý metadata META\_DATH.



### c. Kết quả

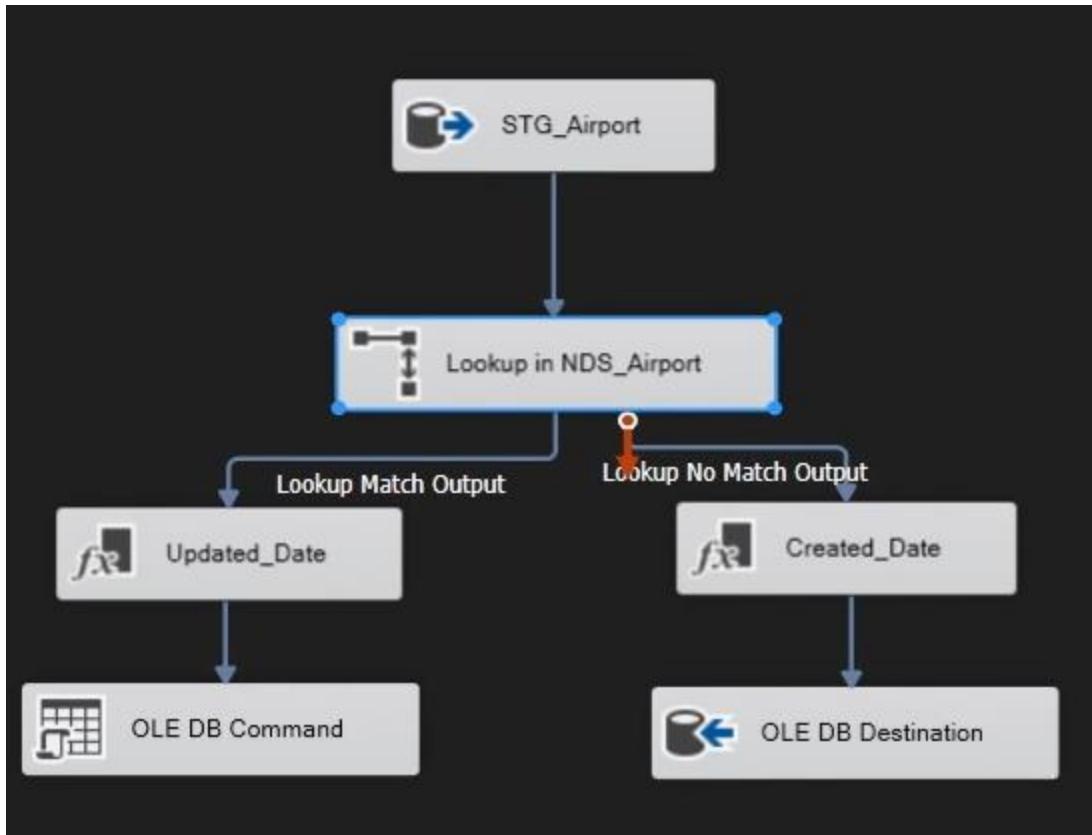
Package Source -> Stage:



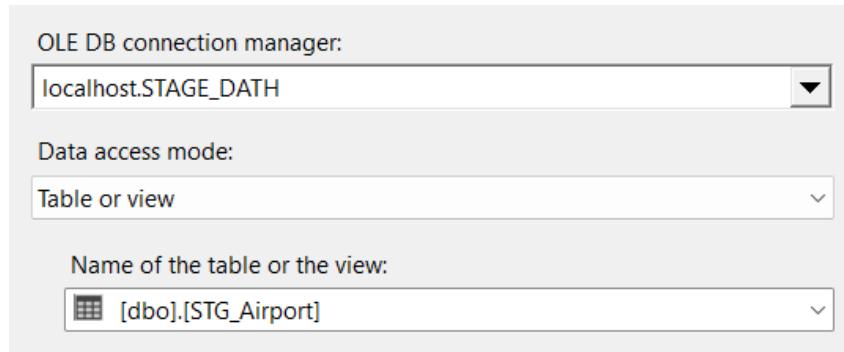
## 2.2. Stage -> NDS



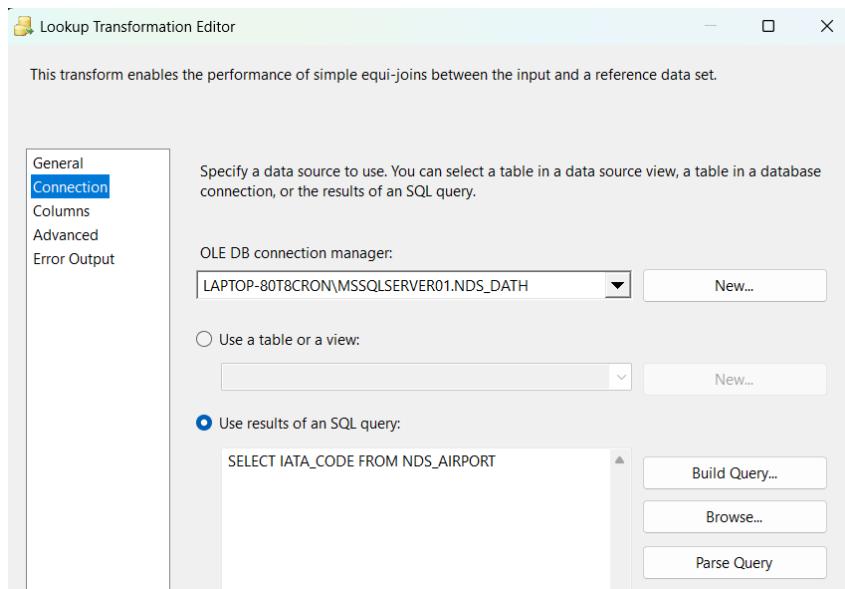
a. Airport



**Bước 1:** Tạo kết nối với bảng STG\_Airport ở Stage.



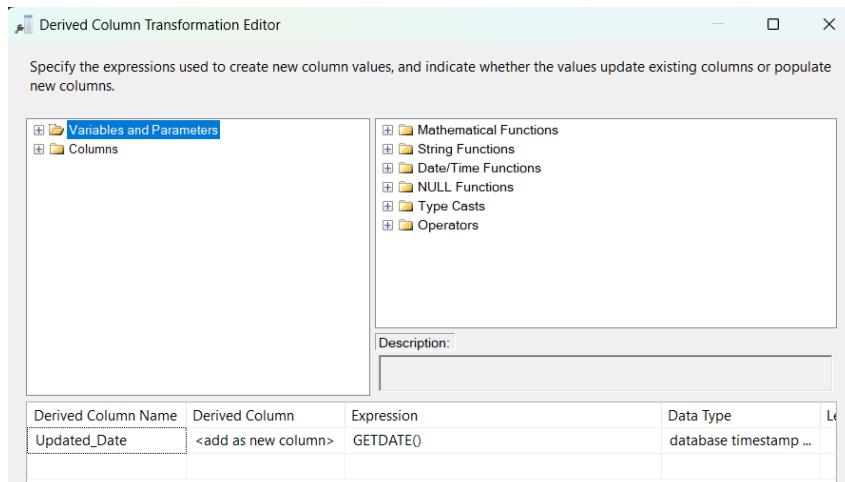
**Bước 2:** Lookup để lấy IATA\_CODE trong bảng NDS\_Airport.



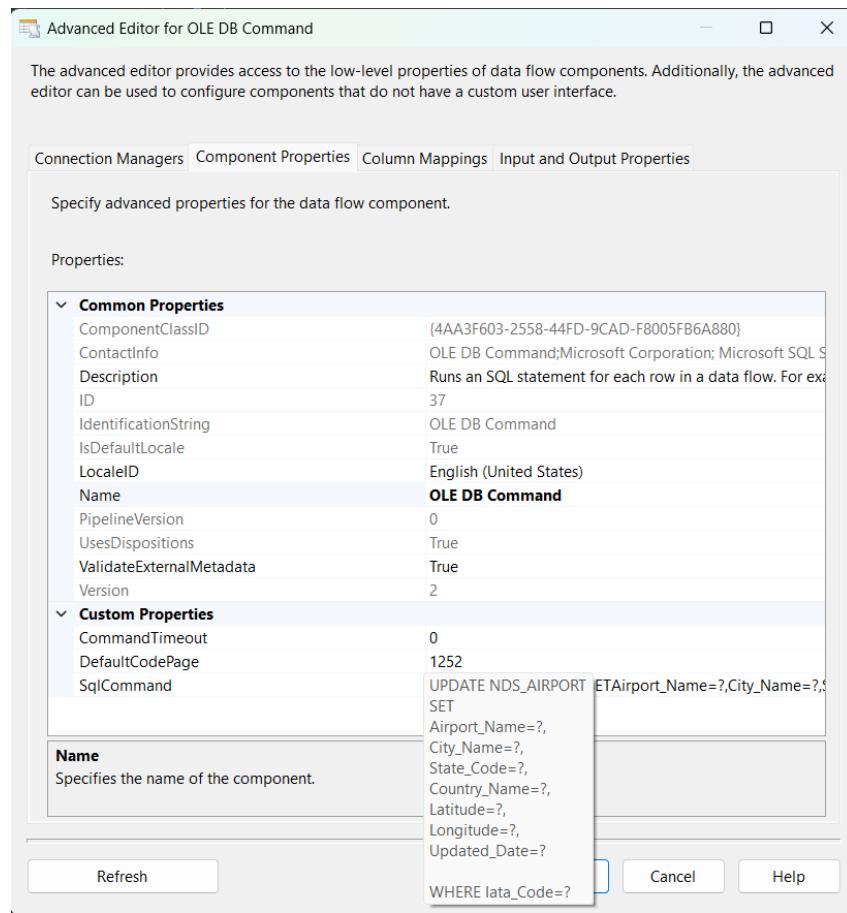
### Bước 3:

#### - Bước 3.1:

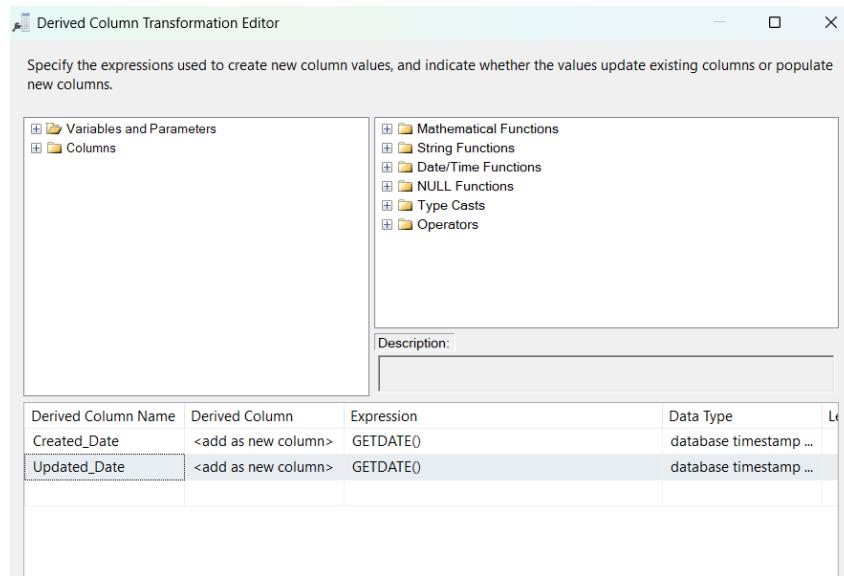
- o Nếu Lookup Match Output (đã có IATA\_CODE trong bảng NDS\_Airport) thì sẽ thực hiện update Updated\_Date thành GetDate().



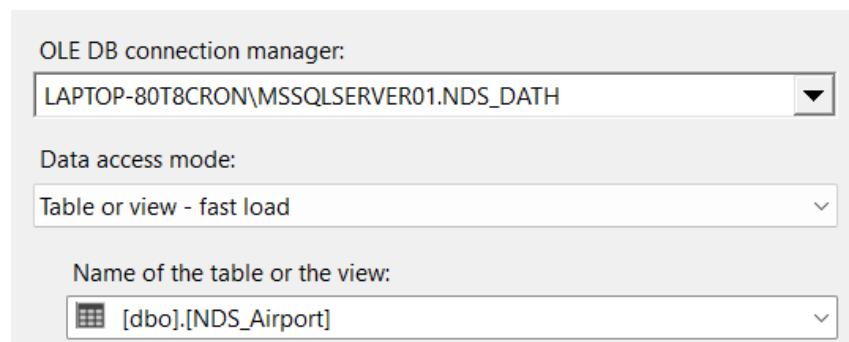
- o Sau đó, cập nhật các thuộc tính có thay đổi thành giá trị mới trong bảng NDS\_Airport.



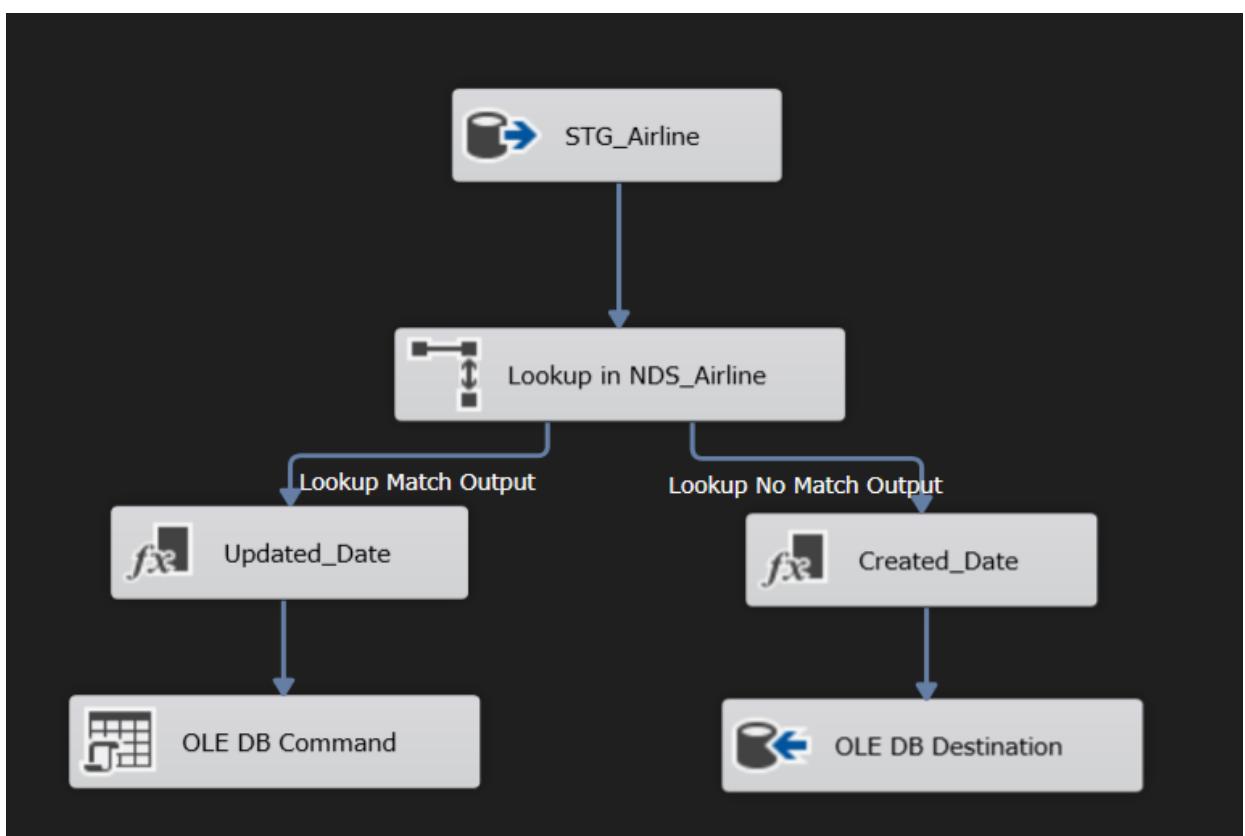
- Bước 3.2:
  - o Nếu Lookup No Match Output (chưa có IATA\_CODE trong bảng NDS\_Airport) thì sẽ thực hiện update Created\_Date và Updated\_Date thành GetDate().



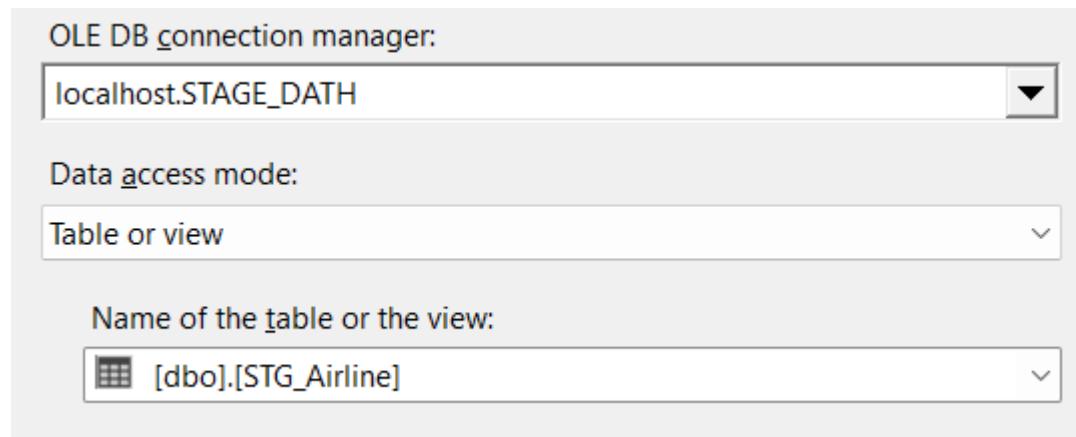
- Sau đó, chọn bảng NDS\_Airport trong để lưu dữ liệu mới xuống.



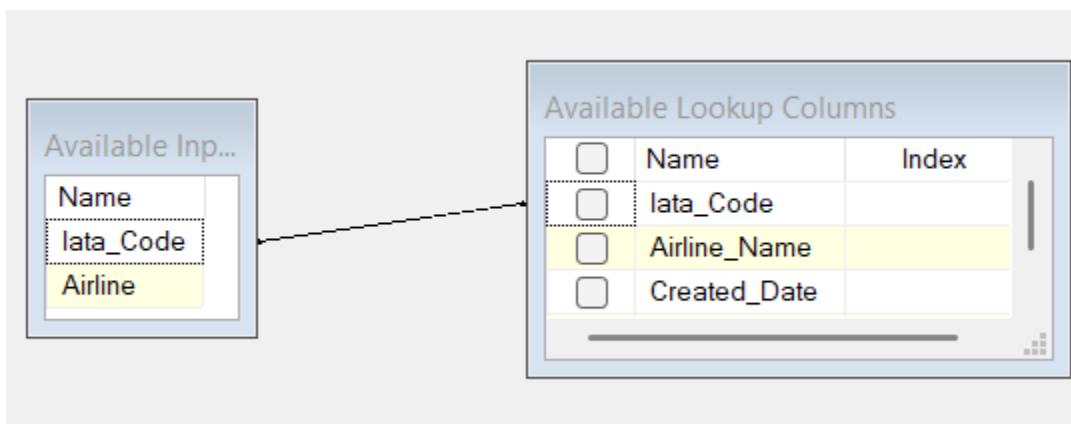
## b. Airline



**Bước 1:** Tạo kết nối với bảng STG\_Airline ở Stage.

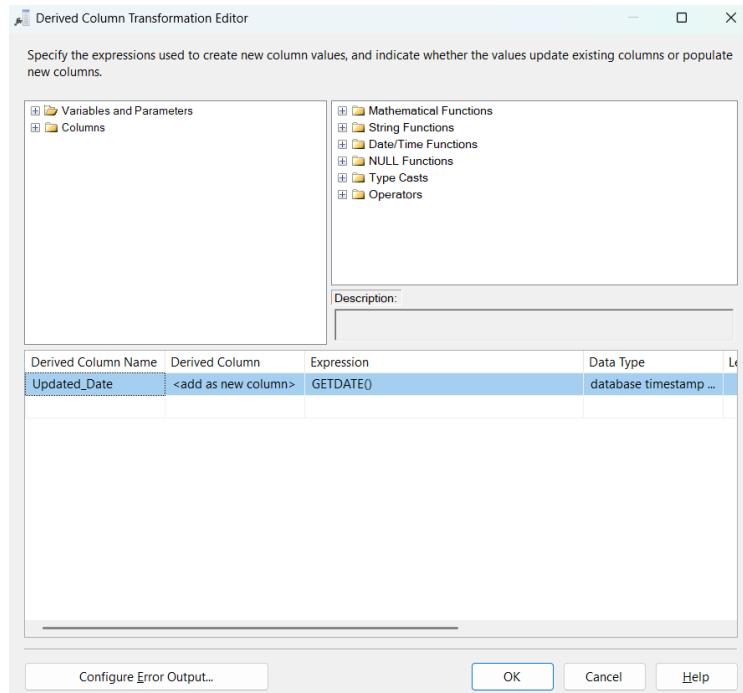


**Bước 2:** Lookup để lấy IATA\_CODE trong bảng NDS\_Airline.

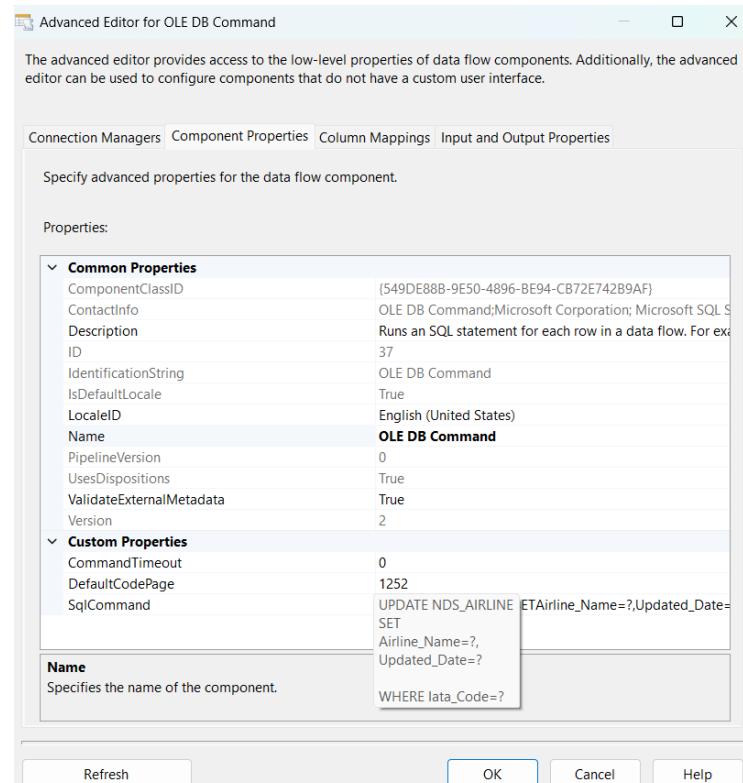


**Bước 3:**

- **Bước 3.1:**
  - o Nếu Lookup Match Output (đã có IATA\_CODE trong bảng NDS\_Airline) thì sẽ thực hiện update Updated\_Date thành GetDate().

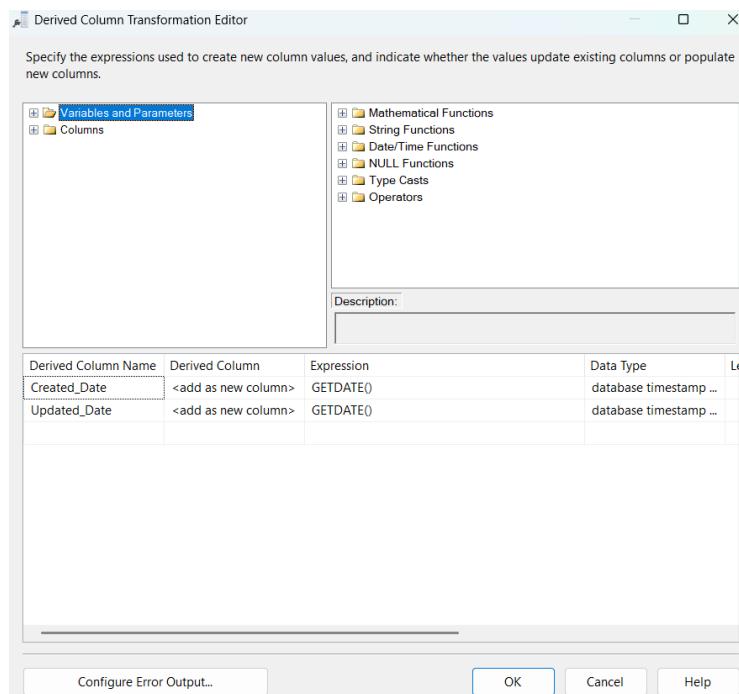


- Sau đó, cập nhật các thuộc tính có thay đổi thành giá trị mới trong bảng NDS\_Airport.

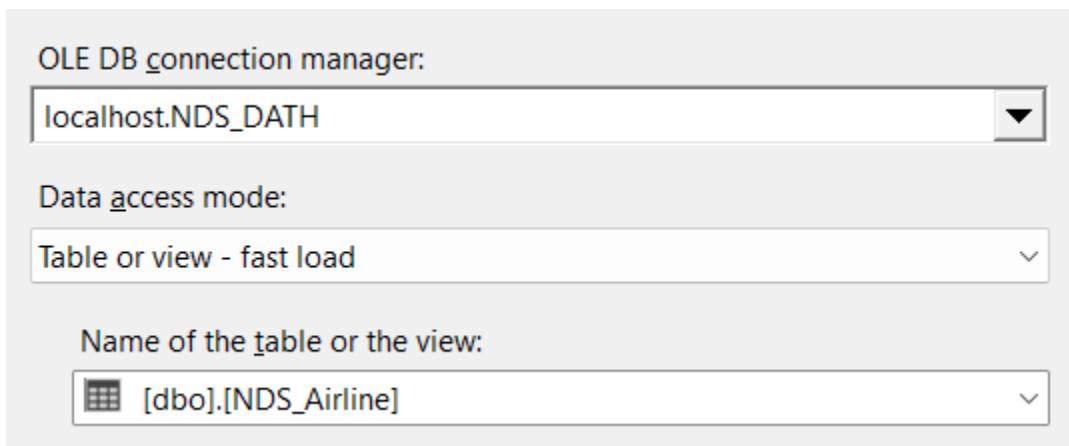


### - Bước 3.2:

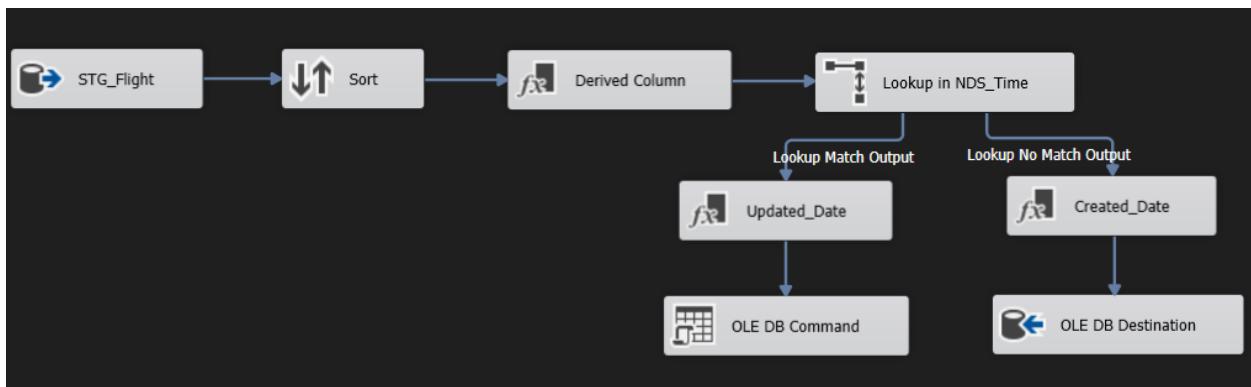
- Nếu Lookup No Match Output (chưa có IATA\_CODE trong bảng NDS\_Airline) thì sẽ thực hiện update Created\_Date và Updated\_Date thành GetDate().



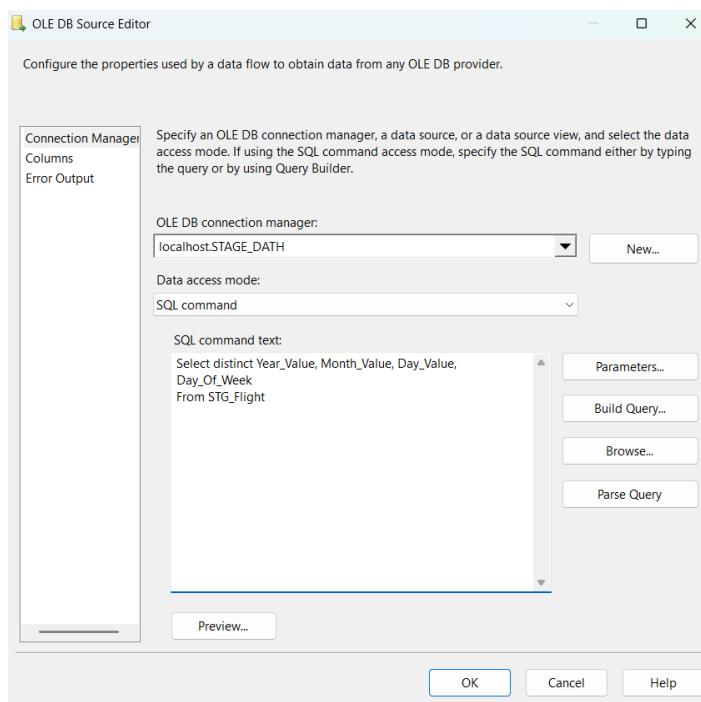
- Sau đó, chọn bảng NDS\_Airline trong để lưu dữ liệu mới xuống.



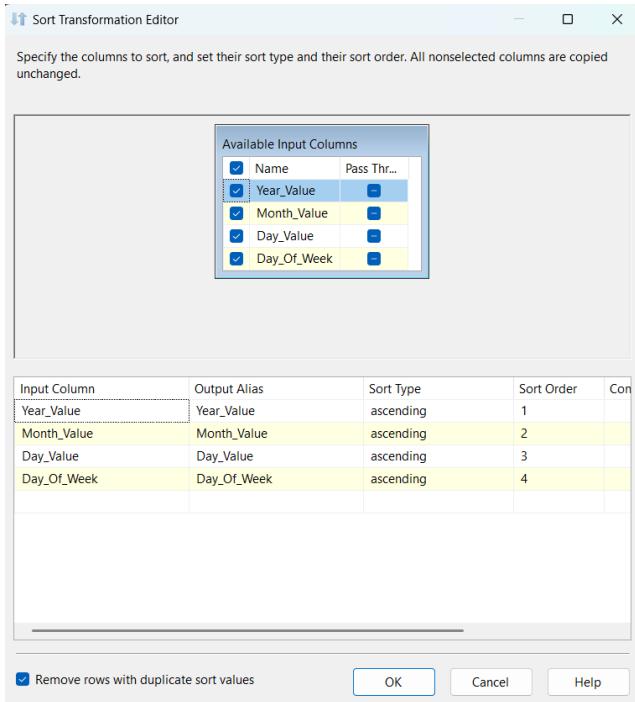
### c. Time



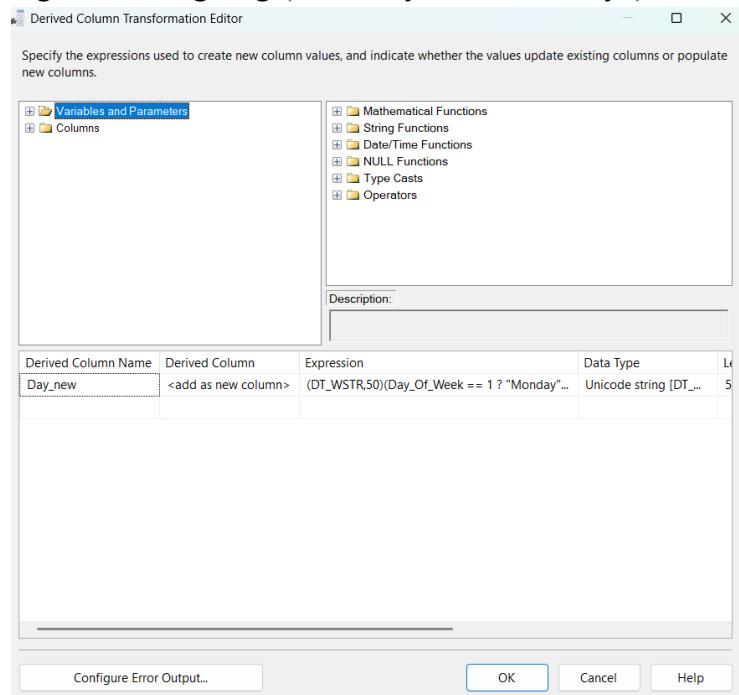
**Bước 1:** Kết nối STG\_Flight và lấy các thuộc tính về thời gian.



**Bước 2:** Thực hiện sắp xếp các cột thời gian trước khi look-up



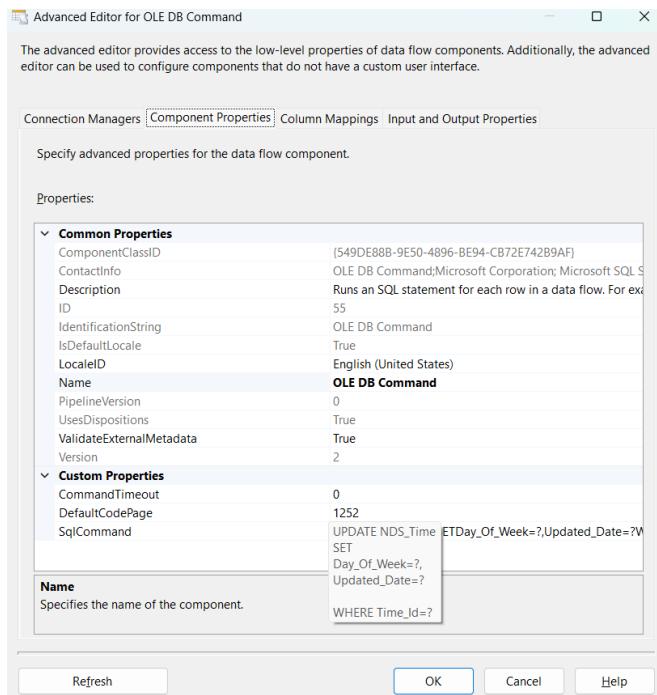
**Bước 3:** Sử dụng Derived Column chuyển đổi mã số ngày trong tuần (1-7) thành tên ngày bằng tiếng Anh tương ứng ("Monday" đến "Sunday").



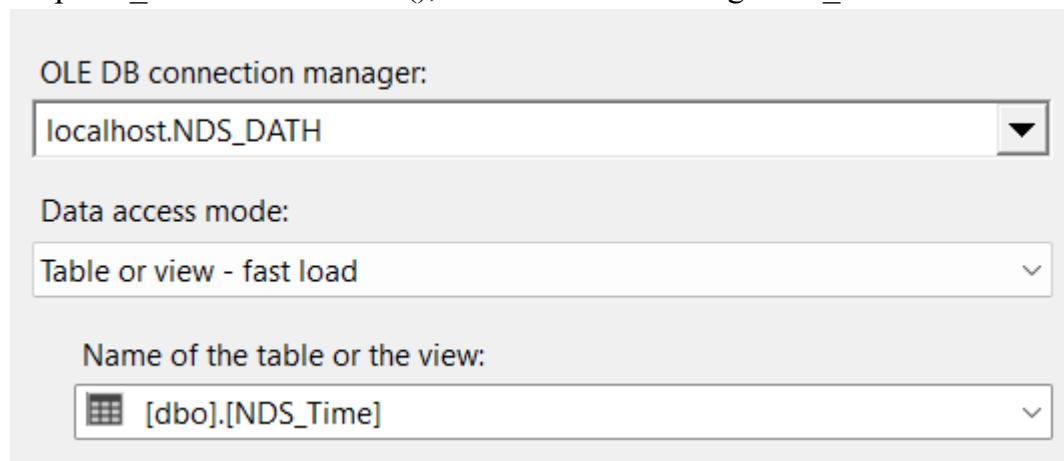
**Bước 4:** Thực hiện Look-up dựa trên các thuộc tính thời gian Year\_Value, Month\_Value, Day\_Value, và Day\_Of\_Week để lấy về giá trị Time\_ID tương ứng.

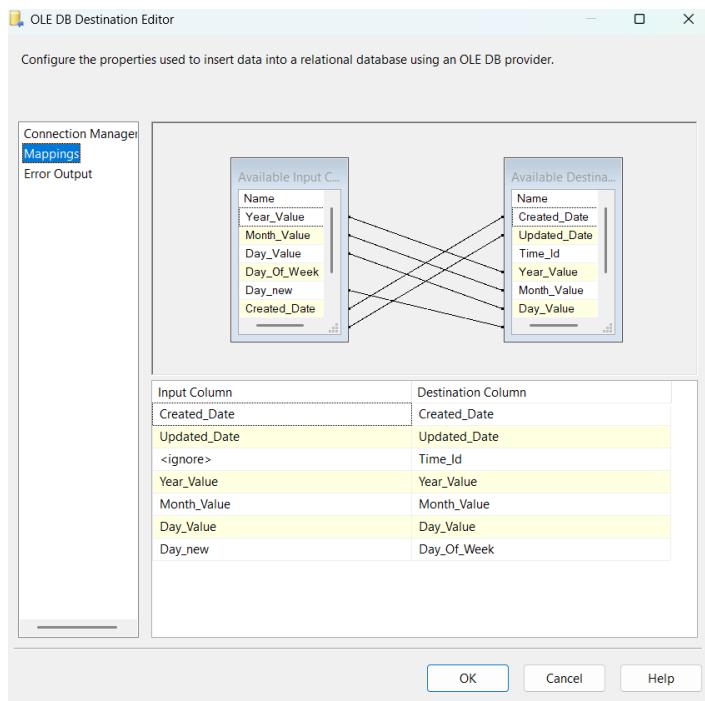
**Bước 5:**

- **Bước 5.1:** Nếu đã tồn tại giá trị Time\_ID, thực hiện Update\_Date = GETDATE() và thực hiện các cập nhật tương ứng vào NDS\_Time.



- **Bước 5.2:** Nếu chưa tồn tại giá trị Time\_ID, tạo mới bản ghi với Created\_Date và Update\_Date = GETDATE(), sau đó chèn vào bảng NDS\_Time.

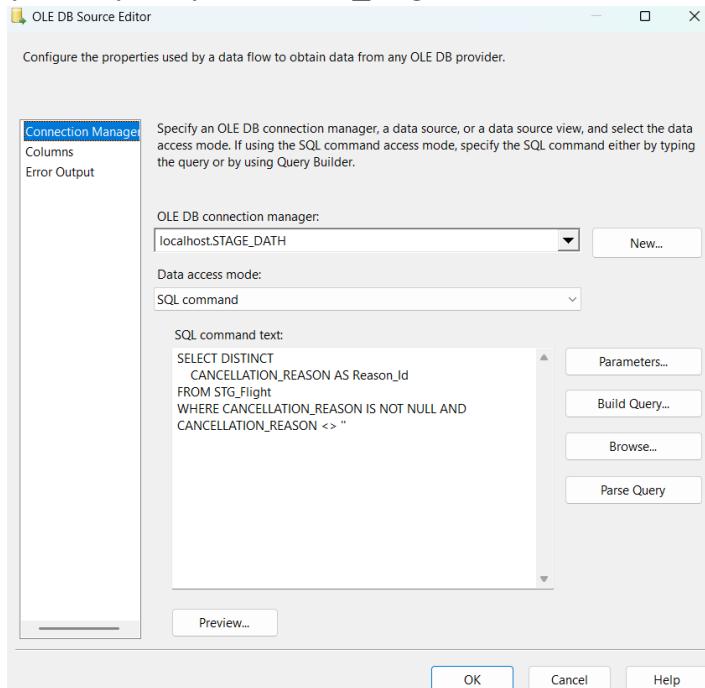




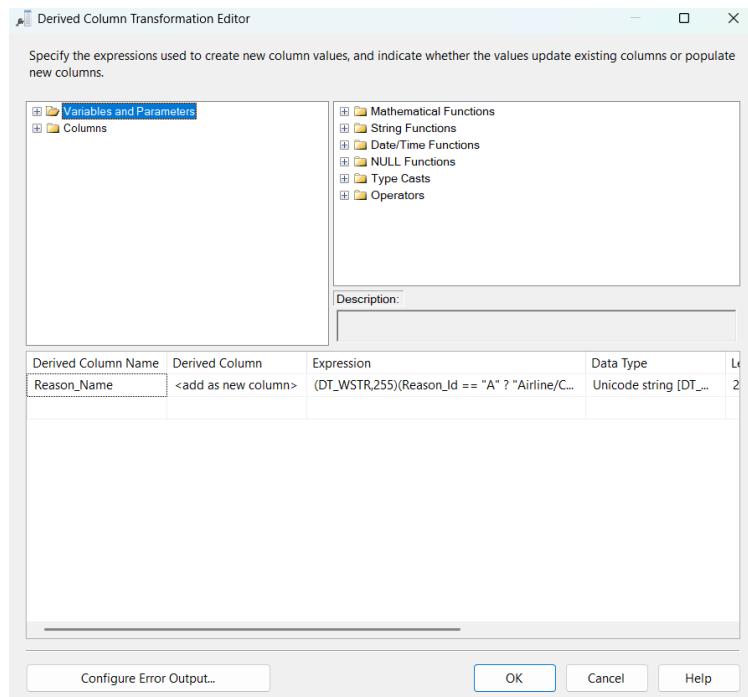
#### d. Reason



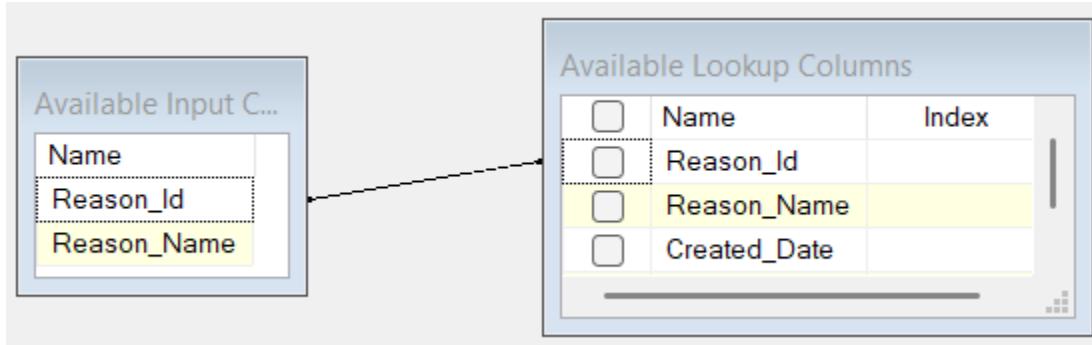
**Bước 1:** Lấy các lý do hủy chuyến từ STG\_Flight.



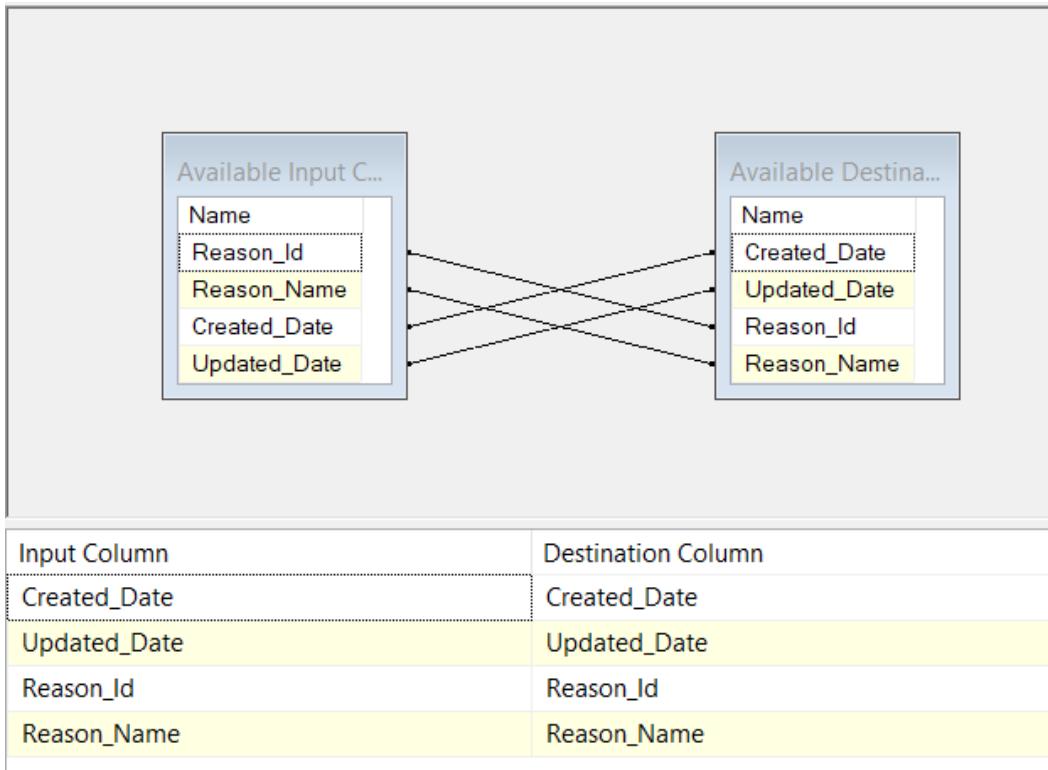
**Bước 2:** Sử dụng Derived chuyển mã lý do hủy chuyến (Reason\_Id) sang mô tả.



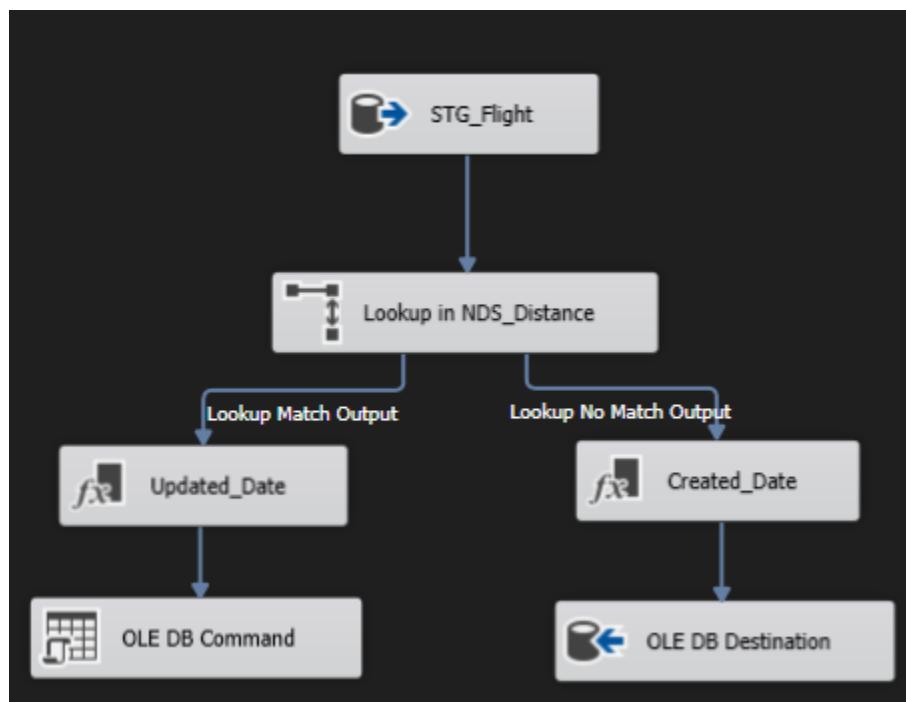
**Bước 3:** Thực hiện tra cứu (Look-up) Reason\_ID trong bảng NDS\_Reason.



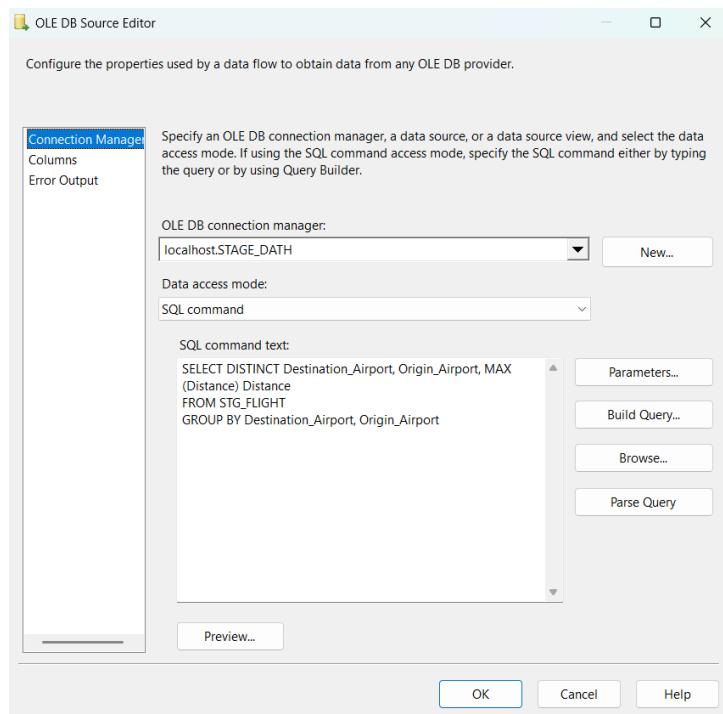
**Bước 4:** Nếu Reason\_ID chưa có trong NDS\_Reason, lấy Created\_Date và Updated\_Date bằng thời gian hiện tại (GETDATE()) rồi thêm bản ghi mới vào bảng.



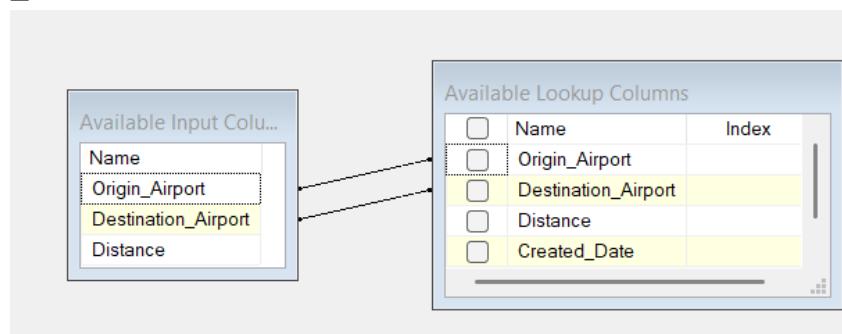
### e. Distance



**Bước 1:** Lấy danh sách duy nhất các cặp sân bay đến (Destination\_Airport) và sân bay đi (Origin\_Airport) từ bảng STG\_FLIGHT, đồng thời lấy giá trị khoảng cách (Distance) lớn nhất tương ứng cho từng cặp đó.

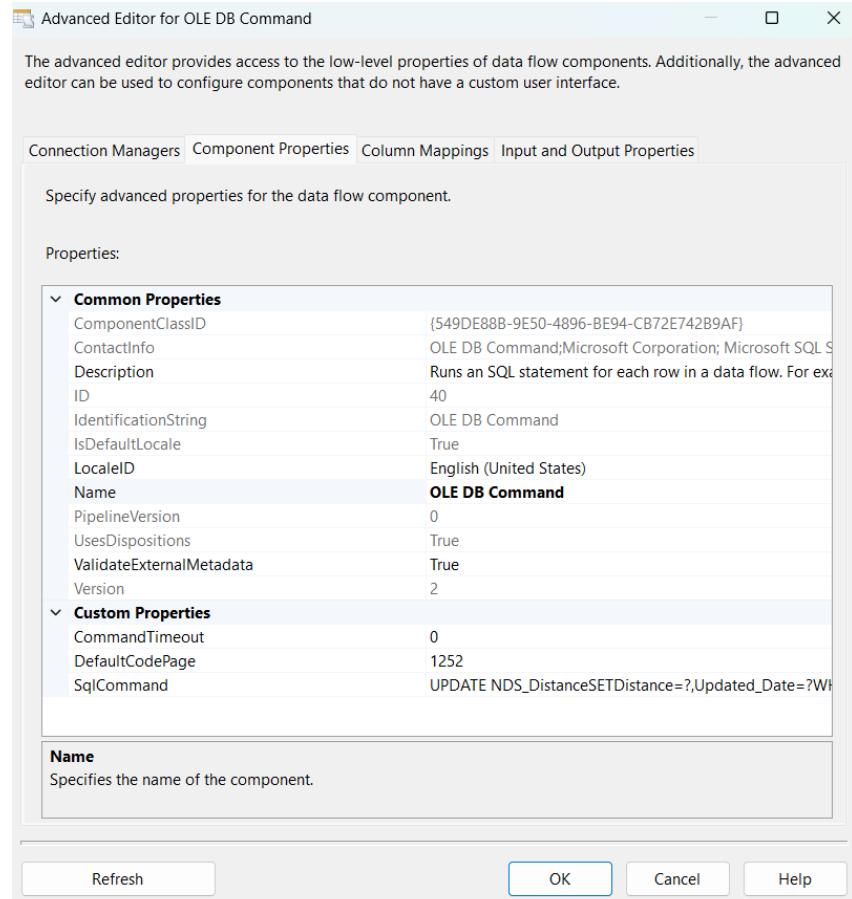


**Bước 2:** Tra cứu (Lookup) các cặp Origin\_Airport và Destination\_Airport trong bảng NDS\_Distance.



**Bước 3:**

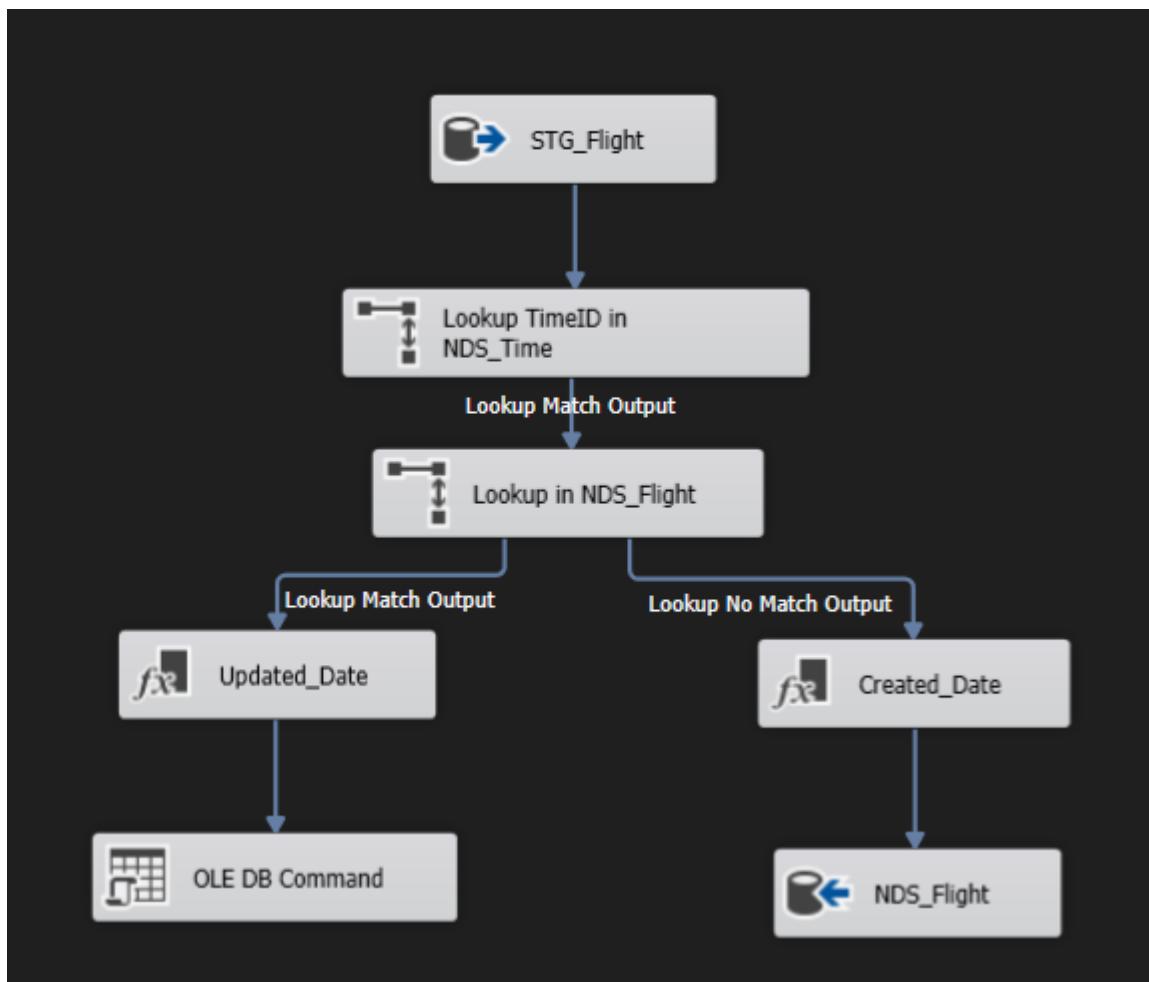
- **Bước 3.1:** Nếu Lookup trùng (Match Output), cập nhật trường Updated\_Date = GETDATE() và các thuộc tính thay đổi trong bảng NDS\_Distance.



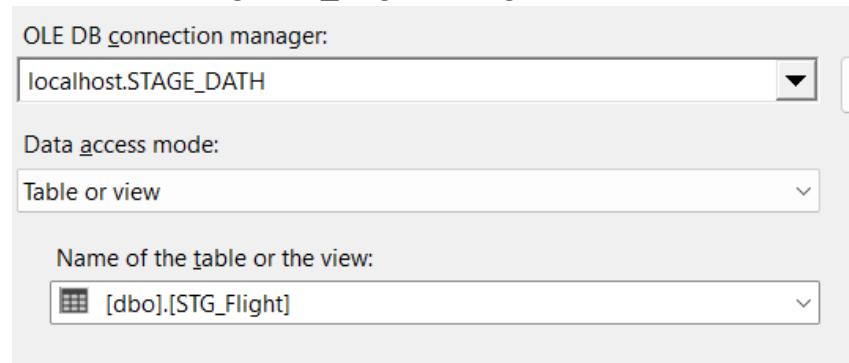
- **Bước 3.2:** Nếu Lookup không trùng (No Match Output), tạo mới bản ghi với Created\_Date và Updated\_Date = GETDATE(), rồi thêm vào bảng NDS\_Distance.



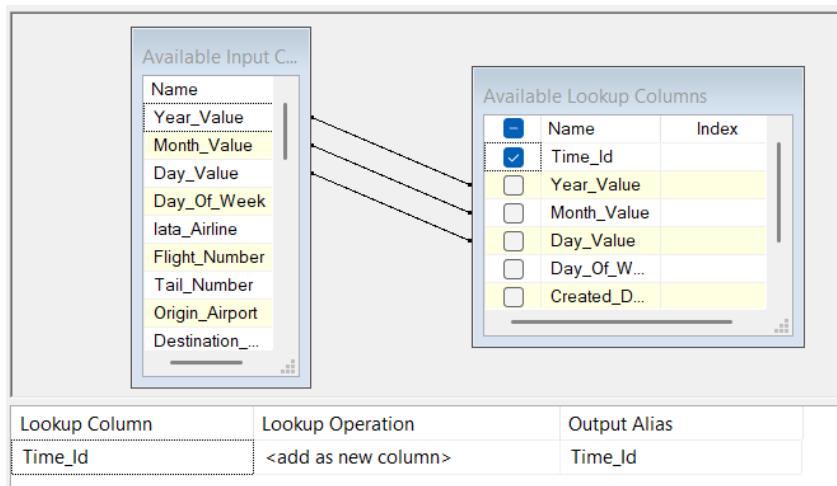
## f. Flight



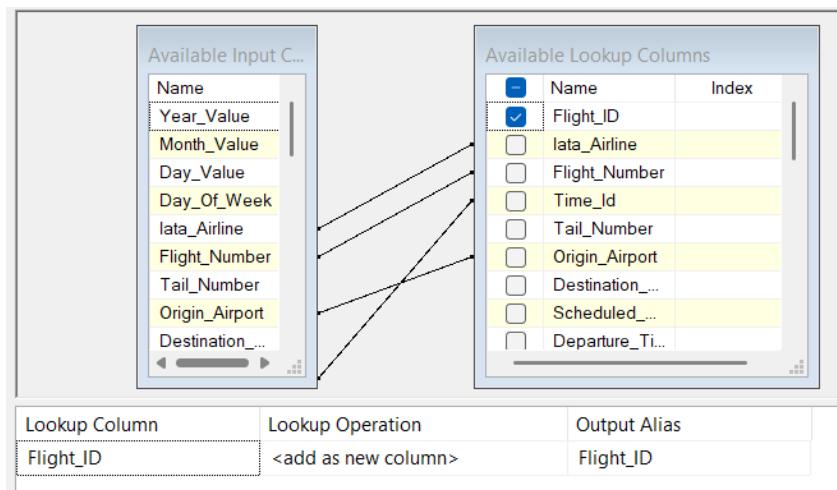
**Bước 1:** Tạo kết nối với bảng STG\_Flight ở Stage.



**Bước 2:** Tra cứu (Lookup) các bộ Year\_Value, Month\_Value, Day\_Value, Day\_Of\_Week trong NDS\_Time và trả về TIME\_ID.

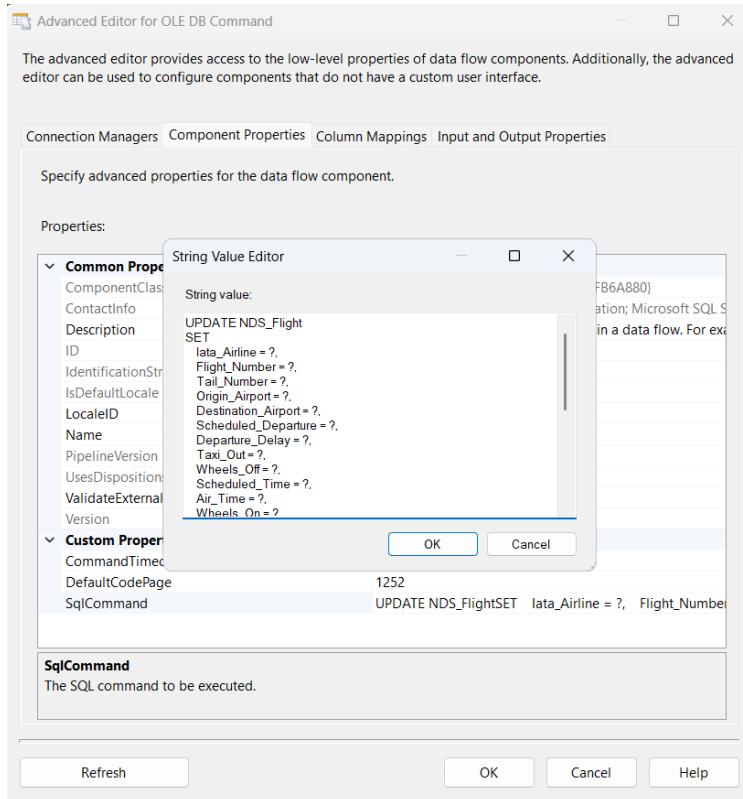


**Bước 3:** Nếu Lookup trùng (Match Output), tiếp tục thực hiện Look-up trên các thuộc tính IATA\_AIRLINE, FLIGHT\_NUMBER, TIME\_ID từ NDS\_Flight và trả về Flight\_ID.

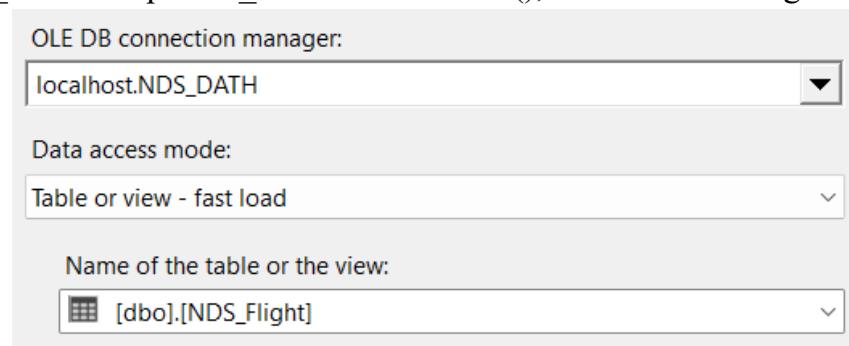


**Bước 4:**

- **Bước 4.1:** Nếu Lookup trùng (Match Output), cập nhật trường Updated\_Date = GETDATE() và các thuộc tính thay đổi trong bảng NDS\_Flight.

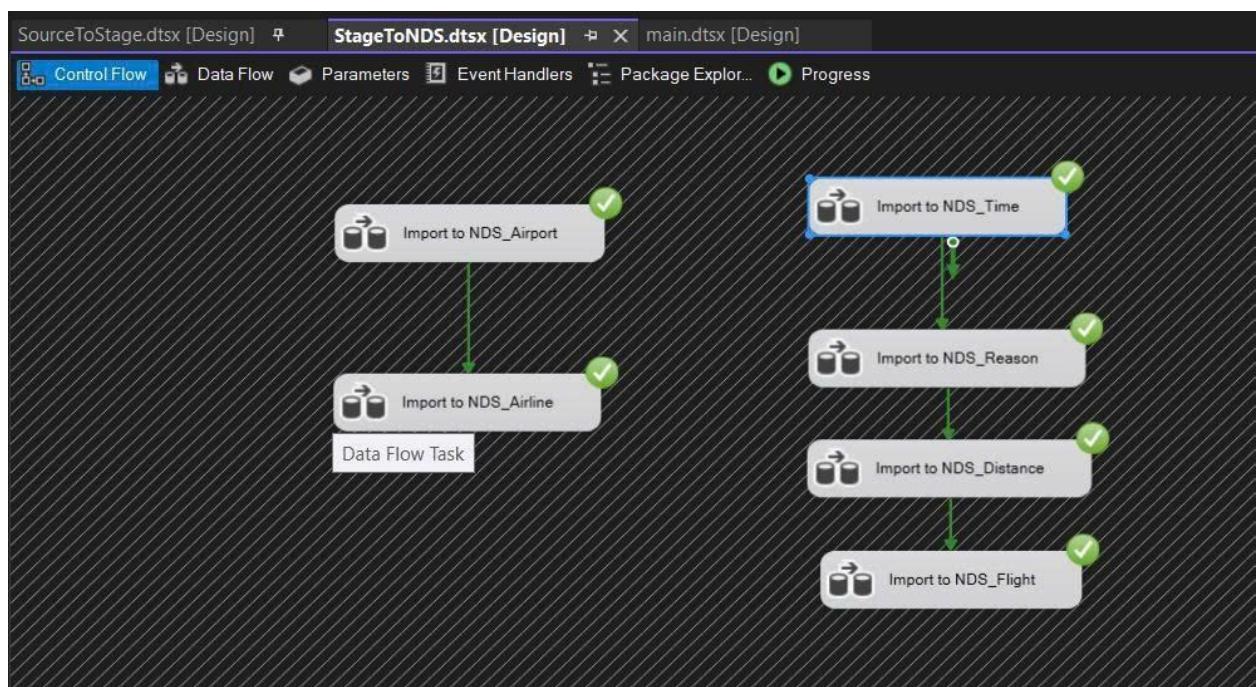


- **Bước 4.2:** Nếu Lookup không trùng (No Match Output), tạo mới bản ghi với Created\_Date và Updated\_Date = GETDATE(), rồi thêm vào bảng NDS\_Flight.



### g. Kết quả

Package Stage -> NDS:



NDS\_Time:

```

48 | SELECT * FROM NDS_Time
49 | SELECT * FROM NDS_Reason
50 | SELECT * FROM NDS_Flight
51 | SELECT * FROM NDS_Distance
52 |

```

123 % ▾

Results Messages

	Time_Id	Year_Value	Month_Value	Day_Value	Day_Of_Week	Created_Date	Updated_Date
1	1	2015	1	1	Thursday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
2	2	2015	1	2	Friday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
3	3	2015	1	3	Saturday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
4	4	2015	1	4	Sunday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
5	5	2015	1	5	Monday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
6	6	2015	1	6	Tuesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
7	7	2015	1	7	Wednesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
8	8	2015	1	8	Thursday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
9	9	2015	1	9	Friday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
10	10	2015	1	10	Saturday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
11	11	2015	1	11	Sunday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
12	12	2015	1	12	Monday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
13	13	2015	1	13	Tuesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
14	14	2015	1	14	Wednesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
15	15	2015	1	15	Thursday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
16	16	2015	1	16	Friday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
17	17	2015	1	17	Saturday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
18	18	2015	1	18	Sunday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
19	19	2015	1	19	Monday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
20	20	2015	1	20	Tuesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
21	21	2015	1	21	Wednesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
22	22	2015	1	22	Thursday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
23	23	2015	1	23	Friday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
24	24	2015	1	24	Saturday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
25	25	2015	1	25	Sunday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
26	26	2015	1	26	Monday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607
27	27	2015	1	27	Tuesday	2025-12-02 20:07:45.607	2025-12-02 20:07:45.607

### NDS\_Reason:

```

49 | SELECT * FROM NDS_Reason
50 | SELECT * FROM NDS_Flight
51 | SELECT * FROM NDS_Distance
52 |

```

123 % ▾

Results Messages

	Reason_Id	Reason_Name	Created_Date	Updated_Date
1	A	Airline/Carrier	2025-12-02 20:07:46.053	2025-12-02 20:07:46.053
2	B	Weather	2025-12-02 20:07:46.053	2025-12-02 20:07:46.053
3	C	National Air System	2025-12-02 20:07:46.053	2025-12-02 20:07:46.053

## NDS\_Flight:

	Flight_ID	Iata_Airline	Flight_Number	Time_Id	Tail_Number	Origin_Airport	Destination_Airport	Scheduled_Departure	Departure_Delay	Taxi_Out	Wheels_Off	Scheduled_Time	Air_Time	Wheels_On	Taxi_In	Scheduled_Arr
1	1	AA	1092	123	N3LDAA	ORD	LAX	2020	-5	54	21:09:00.000000	280	236	23:05:00.000000	7	23:00:00.000000
2	2	AA	2292	123	N3KHAA	DFW	LAS	2025	29	12	21:06:00.000000	187	152	21:38:00.000000	7	21:32:00.000000
3	3	AA	1316	123	N859AA	DFW	MIA	2030	0	17	20:47:00.000000	172	151	00:18:00.000000	6	00:22:00.000000
4	4	AA	2377	123	N473AA	DFW	ORD	2035	16	12	21:03:00.000000	152	106	22:49:00.000000	10	23:07:00.000000
5	5	AA	102	123	N381AA	HNL	DFW	2040	-6	18	20:52:00.000000	457	409	08:41:00.000000	11	09:17:00.000000
6	6	WN	21	123	N614SW	IND	DEN	2045	3	9	20:57:00.000000	165	151	21:28:00.000000	6	21:30:00.000000
7	7	EV	4501	123	N14923	ORD	RIC	2050	23	21	21:34:00.000000	117	93	00:07:00.000000	3	23:47:00.000000
8	8	UA	481	123	N427UA	ORD	BOS	2057	43	14	21:54:00.000000	137	109	00:43:00.000000	6	00:14:00.000000
9	9	WN	3747	123	N405VN	SMF	SNA	2100	-1	10	21:09:00.000000	85	70	22:19:00.000000	19	22:25:00.000000
10	10	EV	5060	123	N860AS	ATL	AGS	2106	62	23	22:31:00.000000	48	28	22:59:00.000000	3	21:54:00.000000
11	11	OO	5256	123	N776SK	IAH	MFE	2110	-3	22	21:29:00.000000	76	47	22:16:00.000000	6	22:26:00.000000
12	12	EV	4862	123	N14125	IAH	AEX	2118	-4	16	21:30:00.000000	60	32	22:02:00.000000	3	22:18:00.000000
13	13	WN	882	123	N603SW	STL	MSP	2125	23	8	21:56:00.000000	90	77	23:13:00.000000	4	22:55:00.000000
14	14	AS	461	123	N760AS	SEA	BLI	2135	11	15	22:01:00.000000	43	21	22:22:00.000000	2	22:18:00.000000
15	15	WN	292	123	N247VN	MDW	ROC	2140	-4	15	21:51:00.000000	90	69	00:00:00.000000	7	00:10:00.000000
16	16	MQ	2910	123	N925MQ	ORD	TVC	2145	2	15	22:02:00.000000	67	36	23:38:00.000000	4	23:52:00.000000
17	17	HA	237	123	N487HA	KOA	HNL	2150	-9	7	21:48:00.000000	43	29	22:17:00.000000	8	22:33:00.000000
18	18	EV	5053	123	N752EV	LGA	ROC	2155	-5	26	22:16:00.000000	79	43	22:59:00.000000	6	23:14:00.000000
19	19	MQ	3108	123	N657MQ	DFW	TYR	2200	5	15	22:20:00.000000	49	21	22:41:00.000000	6	22:49:00.000000

## NDS\_Distance:

51 | SELECT \* FROM NDS\_Distance  
52

123 %

Results Messages

	Origin_Airport	Destination_Airport	Distance	Created_Date	Updated_Date
1	ABE	ATL	692	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
2	ABE	DTW	425	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
3	ABE	ORD	655	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
4	ABI	DFW	158	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
5	ABQ	ATL	1269	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
6	ABQ	BWI	1670	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
7	ABQ	DAL	580	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
8	ABQ	DEN	349	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
9	ABQ	DFW	569	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
10	ABQ	HOU	759	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
11	ABQ	IAH	744	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
12	ABQ	JFK	1826	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
13	ABQ	LAS	486	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
14	ABQ	LAX	677	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
15	ABQ	MCI	718	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
16	ABQ	MCO	1553	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
17	ABQ	MDW	1121	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
18	ABQ	MSP	981	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
19	ABQ	OAK	889	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
20	ABQ	ORD	1118	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
21	ABQ	PHX	328	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323
22	ABQ	SAN	628	2025-12-02 20:07:46.323	2025-12-02 20:07:46.323

## NDS -> DDS

### 1. Bảng Mapping

#### 1.1. Mapping NDS\_Airline → Dim\_Airline (SCD Type 2)

NDS	DDS Target	Xử lý trong SSIS	Giải thích
Iata_Code	IATA_Code	Business Key trong SCD	Dùng làm khóa định danh để so sánh dữ liệu cũ và mới.
Airline_Name	Airline_Name	Historical Attribute	SCD tự động phát hiện thay đổi tên để tạo bản ghi mới (SCD Type 2).
—	Is_Active	Derived Column	Gán giá trị 0 (ánh xạ từ (DT_BOOL)("0")) cho bản ghi cũ và 1 (ánh xạ từ (DT_BOOL)("1")) cho bản ghi mới.
—	Airline_Key	Surrogate Key	Khóa chính tự tăng (Identity) trong bảng đích DIM_AIRLINE.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

## 1.2. Mapping NDS\_Airport → Dim\_Airport (SCD Type 2)

NDS Source	DDS Target	Xử lý trong SSIS	Giải thích nghiệp vụ
Iata_Code	IATA_Code	Business Key trong SCD	Dùng làm khóa định danh duy nhất để so sánh dữ liệu.
Airport_Name	Airport_Name	Historical Attribute	SCD tạo bản ghi mới nếu tên sân bay thay đổi (SCD Type 2).
City_Name	City	Historical Attribute	SCD tạo bản ghi mới nếu thành phố thay đổi.
State_Code	State	Historical Attribute	SCD tạo bản ghi mới nếu bang/vùng thay đổi.
Country_Name	Country	Historical Attribute	SCD tạo bản ghi mới nếu quốc gia thay đổi.
Latitude	Latitude	Fixed Attribute	Giữ nguyên giá trị vĩ độ địa lý.

Longitude	Longitude	Fixed Attribute	Giữ nguyên giá trị kinh độ địa lý.
—	Is_Active	Derived Column	Gán 0 cho bản ghi cũ (lịch sử) và 1 cho bản ghi mới nhất.
—	Airport_Key	Surrogate Key	Khóa chính tự tăng (Identity) trong bảng DIM_AIRPORT.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

### 1.3. Mapping NDS\_Reason → Dim\_Reason (SCD Type 1)

NDS Source	DDS Target	Xử lý trong SSIS	Giải thích nghiệp vụ
ReasonID	Reason_Code	Business Key trong SCD	Dùng để định danh các mã lý do (A, B, C, D).
Reason_Name	Reason_Description	Changing Attribute	Ghi đè trực tiếp mô tả mới nhất vào bản ghi cũ (SCD Type 1).
—	Reason_Key	Surrogate Key	Khóa chính tự tăng (Identity) trong bảng DIM_REASON.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

### 1.4. Mapping NDS\_Time → Dim\_Date

NDS Source	DDS Target	Xử lý & Chuẩn hóa (T-SQL Script)	Giải thích
Year, Month, Day	Date_Key	CONVERT(INT, FORMAT(Full_Date, 'yyyyMMdd'))	Tạo khóa chính duy nhất dạng số để tối ưu hóa việc Join dữ liệu.

Year, Month, Day	Full_Date	DATEFROMPARTS(Year, Month, Day)	Chuẩn hóa về kiểu dữ liệu DATE chuẩn.
Year	Year	Direct Map	Phân tích theo năm.
—	Quarter	DATEPART(QUARTER, Full_Date)	Chuẩn hóa: Tự động xác định Quý (1-4) từ ngày.
Month	Month	Direct Map	Phân tích theo tháng số.
—	Month_Name	DATENAME(MONTH, Full_Date)	Lấy tên tháng đầy đủ (January, February...) để hiển thị trên Dashboard.
Day	Day	Direct Map	Ngày trong tháng.
—	Day_Of_Week	DATEPART(WEEKDAY, Full_Date)	Trả về số thứ tự ngày trong tuần (1=Chủ nhật).
Day_Of_Week	Day_Name	DATENAME(WEEKDAY, Full_Date)	Lấy tên thứ (Monday, Tuesday...).
—	Is_Weekend	CASE WHEN DATEPART(WEEKDAY, Full_Date) IN (1, 7) THEN 1 ELSE 0 END	Chuẩn hóa: Phân loại ngày cuối tuần (1) và ngày thường (0).
—	Season	CASE WHEN Month IN (12, 1, 2) THEN 'Winter' ... ELSE 'Autumn' END	Chuẩn hóa: Phân loại mùa (Xuân, Hạ, Thu, Đông) để

			đánh giá ảnh hưởng thời tiết.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

### 1.5. Mapping NDS\_Flight → Dim\_Time\_Of\_Date

NDS Source	DDS Target	Xử lý & Chuẩn hóa (T-SQL Script)	Giải thích nghiệp vụ
Scheduled_Departure	Hour	(DT_I4)(Scheduled_Departure / 100)	Trích xuất phần giờ từ định dạng HHmm (Ví dụ: 1830 -> 18).
—	TimeOfDayName	CASE WHEN @Hour BETWEEN 5 AND 8 THEN 'Early Morning' ... END	Chuẩn hóa: Chia nhóm giờ thành các buổi (Sáng, Trưa, Chiều, Tối, Đêm).
—	TimeOfDayKey	IDENTITY(1,1)	Khóa thay thế (Surrogate Key) tự tăng trong DDS.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

#### 1.5.1. Mapping NDS\_Flight → Fact\_Flight

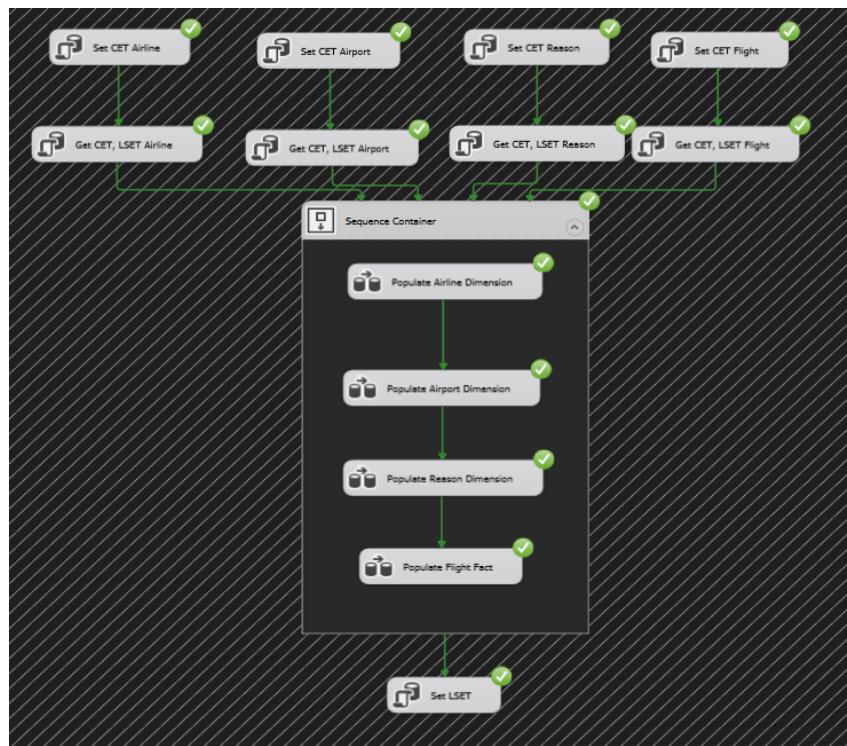
NDS Source	DDS Target	Transformation (Derived Column & Lookup)	Ý nghĩa
NDS_Time.Date_Key	Date_Key	Lookup Dim_Date	Liên kết ngày chuyến bay.

NDS_Flight.Iata_Airline	Airline_Key	Lookup Dim_Airline (Business Key = Iata_Airline)	Liên kết khóa hãng bay hiện hành.
NDS_Flight.Origin_Airport	Origin_Airport_Key	Lookup Dim_Airport (Business Key = Origin_Airport)	Liên kết sân bay xuất phát.
NDS_Flight.Destination_Airport	Dest_Airport_Key	Lookup Dim_Airport (Business Key = Destination_Airport)	Liên kết sân bay đến.
NDS_Flight.Cancellation_Reason	Reason_Key	Lookup Dim_Reason (Mã lý do). Nếu Cancelled = FALSE, gán mặc định khóa của lý do "N".	Liên kết nguyên nhân hủy.
NDS_Flight.Scheduled_Departure	Time_Of_Day_Key	Derived: Scheduled_Departure < 0 ? -1 : (DT_I4)(Scheduled_Departure / 100) → Lookup Dim_TimeOfDay.	Phân tích theo khung giờ bay.
NDS_Flight.Flight_Number	Flight_Number	Direct Map.	Chiều thoái hóa (Degenerate Dimension).
—	Flight_Count	Derived: Gán cố định giá trị 1.	Dùng để đếm tổng số chuyến bay.
NDS_Flight.Departure_Delay	Dep_Delay_Minutes	Derived: ISNULL(Departure_Delay) ? 0 : (DT_I4)Departure_Delay.	Thời gian trễ khởi hành.
NDS_Flight.Arrival_Delay	Arr_Delay_Minutes	Derived: (DT_I4)Arrival_Delay. Nếu NULL gán 0.	Thời gian trễ đến nơi.
NDS_Flight.Cancelled	Is_Cancelled	Derived: Cancelled == TRUE ? 1 : 0.	KPI: Cờ báo hủy chuyến (0/1).

NDS_Flight.Diverted	Is_Diverted	Derived: Diverted == TRUE ? 1 : 0.	KPI: Cờ báo chuyển hướng (0/1).
NDS_Flight.Cancelled, Arrival_Delay	Is OTP	Derived: (Cancelled == TRUE) ? NULL(DT_I4) : (ISNULL(Arrival_Delay) ? 0 : (Arrival_Delay <= 15 ? 1 : 0)).	KPI: Đúng giờ (On-Time Performance).
Air_System_Delay, Security_Delay, Airline_Delay, Late_Aircraft_Delay, Weather_Delay	Air_System_Delay, Security_Delay, Airline_Delay, Late_Aircraft_Delay, Weather_Delay	Derived: ISNULL(Cột) ? 0 : (DT_I4)Cột.	Phân tích chi tiết các loại delay.
—	CREATED_DATE	GETDATE()	Thời điểm nạp bản ghi mới vào DDS.
—	UPDATED_DATE	GETDATE()	Thời điểm cập nhật trạng thái bản ghi.

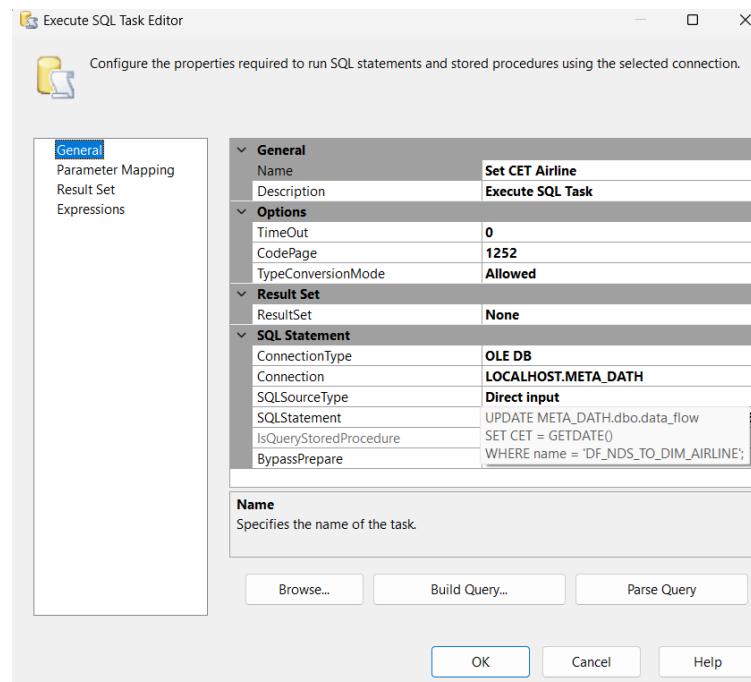
## 2. Giải thích các Control và ETL Component

Mô tả quy trình nạp dữ liệu từ NDS vào DDS

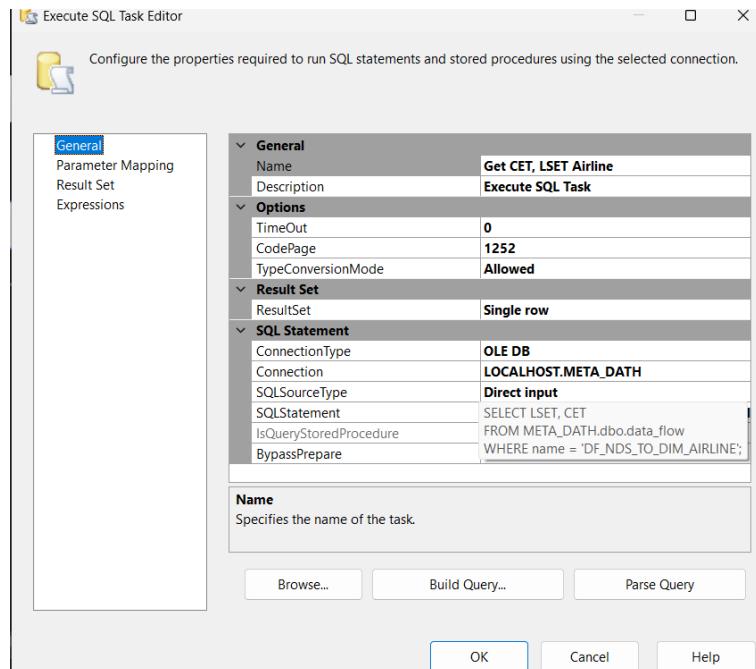


## 2.1. Giai đoạn 1: Chuẩn bị Giá trị CET/LSET

**Bước 1: Set CET Flight:** Cập nhật trường CET (Current Extract Time) của bản ghi quản lý từng bảng trong bảng metadata META\_DATH.data\_flow bằng timestamp hệ thống hiện tại (GETDATE()). Bước này được thực hiện cho tất cả các bảng.



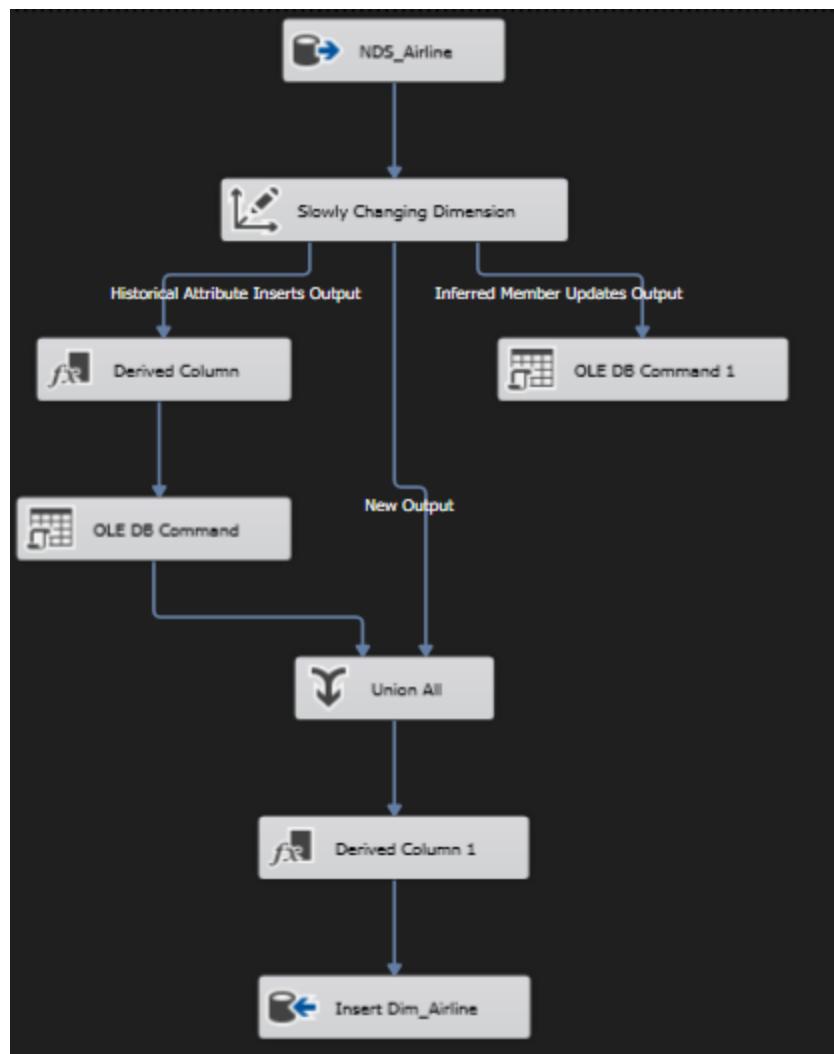
**Bước 2: Get CET, LSET Flight:** Truy xuất thời điểm LSET (Last Successful Extract Time) và CET (Current Extract Time) của các bảng từ CSDL Metadata. Đây là giá trị lần extract thành công gần nhất và thời điểm khởi tạo package hiện tại, dùng để xác định phạm vi dữ liệu mới cần nạp vào DDS. Thực hiện câu lệnh SELECT LSET, CET trên bảng metadata, cấu hình Result Set ở chế độ *Single row*, và gắn kết quả tương ứng vào biến User::LSET và User::CET.



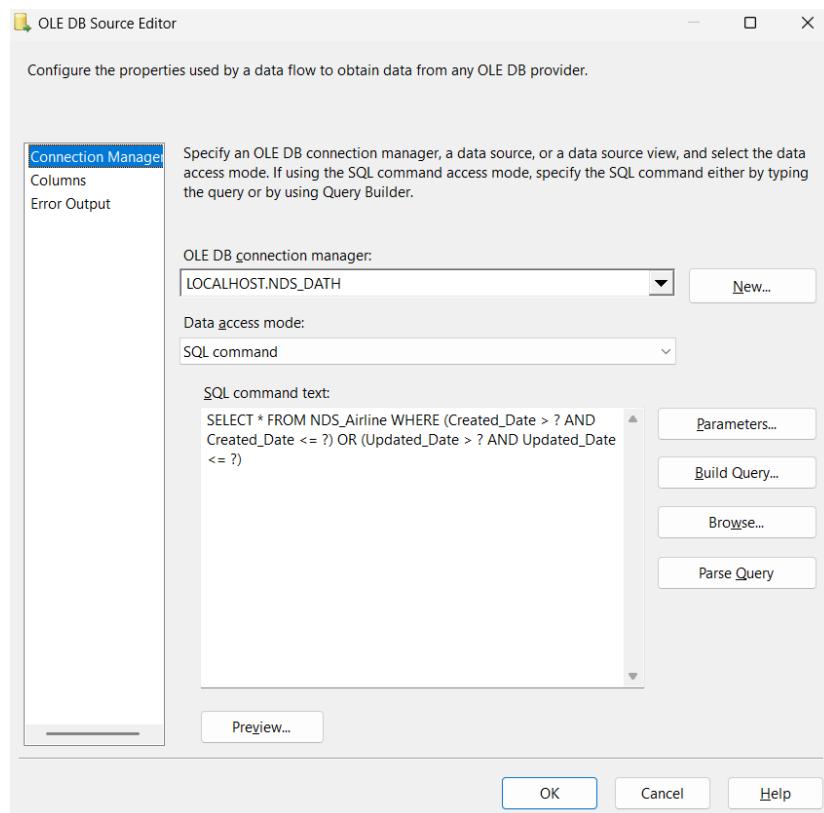
## 2.2. Giai đoạn 2: Nạp dữ liệu

### 2.2.1. Nạp các bảng chiều

#### a. Dim\_Airline



**Bước 1:** Sử dụng OLE DB Source kết nối đến bảng NDS\_Airline. Truy vấn các bản ghi có CreatedDate hoặc UpdatedDate nằm trong khoảng LSET -> CET. Dữ liệu đầu ra gồm các hãng bay mới hoặc đã thay đổi kể từ lần ETL trước.



## Bước 2: Xử lý thay đổi bằng Slowly Changing Dimension (SCD)

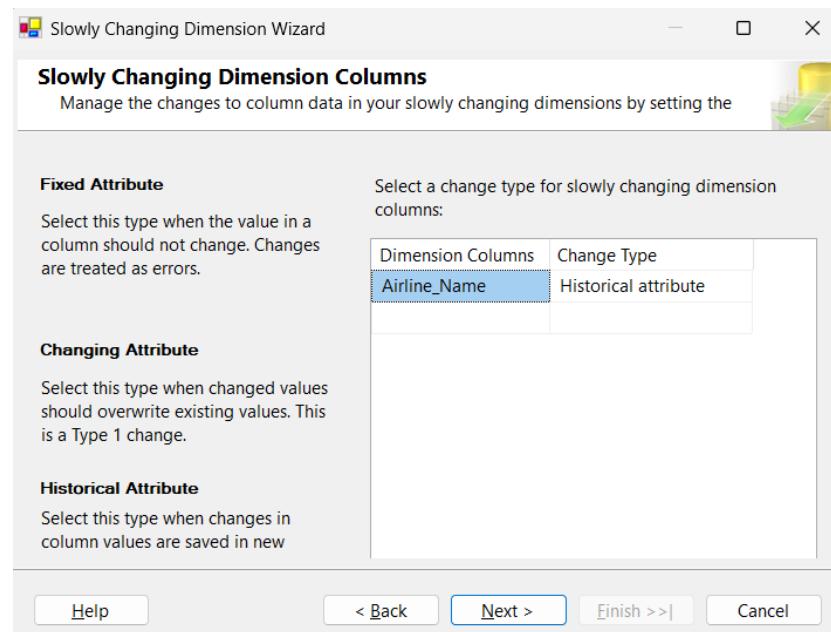
Bước 2.1: Đưa dữ liệu từ NDS vào Slowly Changing Dimension Component.

Bước 2.2: Xác định Business Key là IATA\_Code.

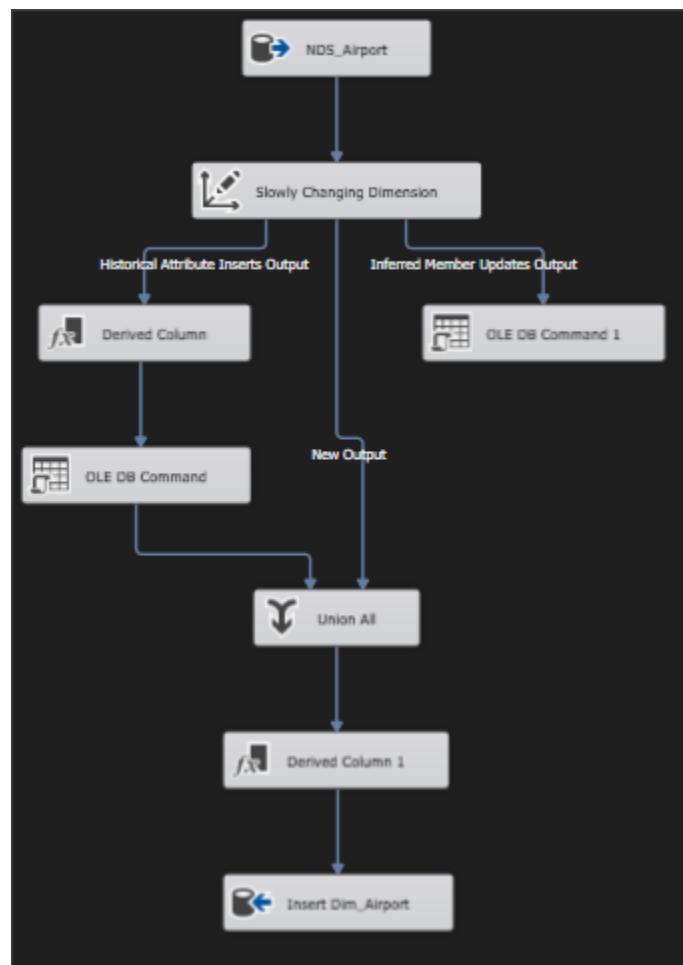
Bước 2.3: Cấu hình Airline\_Name là Historical Attribute (SCD Type 2).

Bước 2.4: Khi Airline\_Name thay đổi:

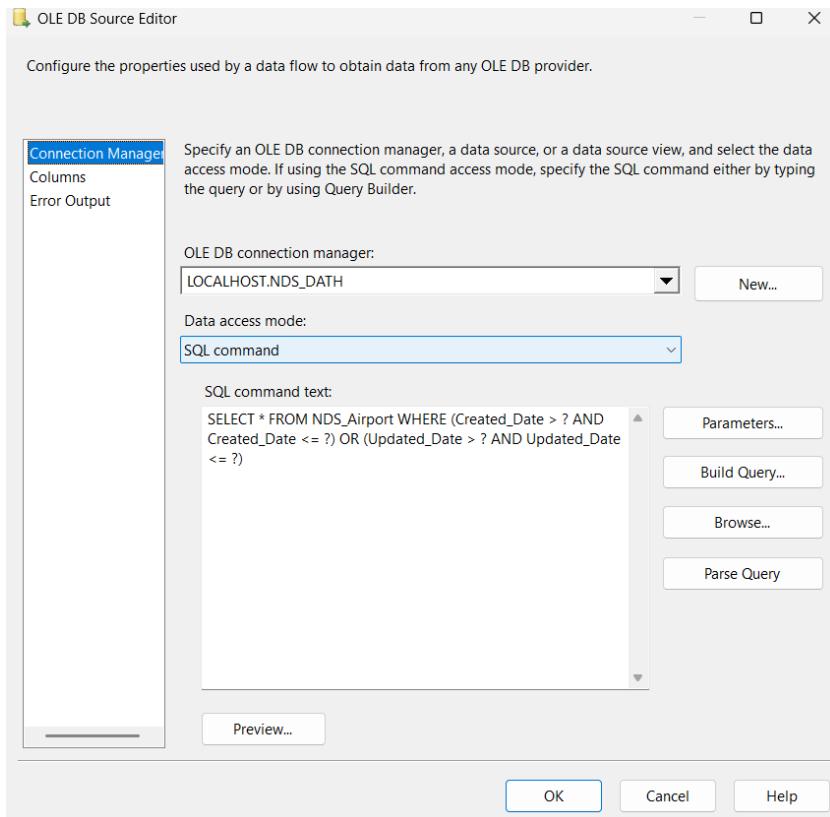
- Cập nhật bản ghi cũ trong Dim\_Airline với Is\_Active = 0.
- Chèn bản ghi mới với khóa thay thế tự tăng và Is\_Active = 1.



## b. Dim\_Airport



**Bước 1:** Sử dụng OLE DB Source kết nối đến bảng NDS\_Airport. Truy vấn các bản ghi có CreatedDate hoặc UpdatedDate nằm trong khoảng LSET -> CET. Dữ liệu đầu ra gồm các sân bay mới hoặc đã thay đổi kể từ lần ETL trước.



## **Bước 2: Xử lý thay đổi bằng Slowly Changing Dimension (SCD)**

Bước 2.1: Đưa dữ liệu từ NDS vào Slowly Changing Dimension Component.

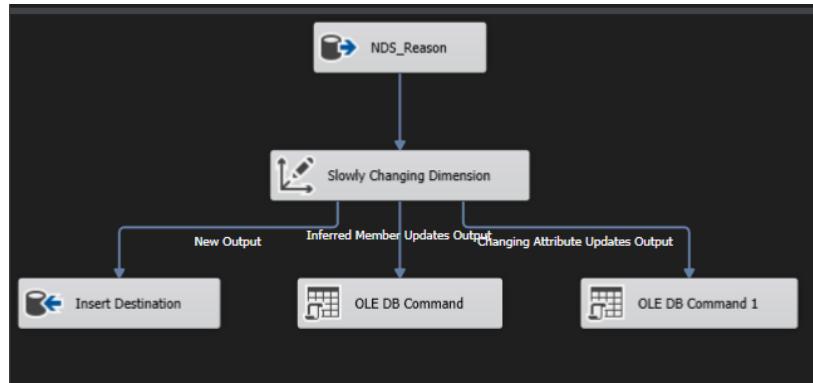
Bước 2.2: Xác định Business Key là IATA\_Code.

Bước 2.3: Cấu hình các thuộc tính Airport\_Name, City, State, Country là Historical Attribute (SCD Type 2).

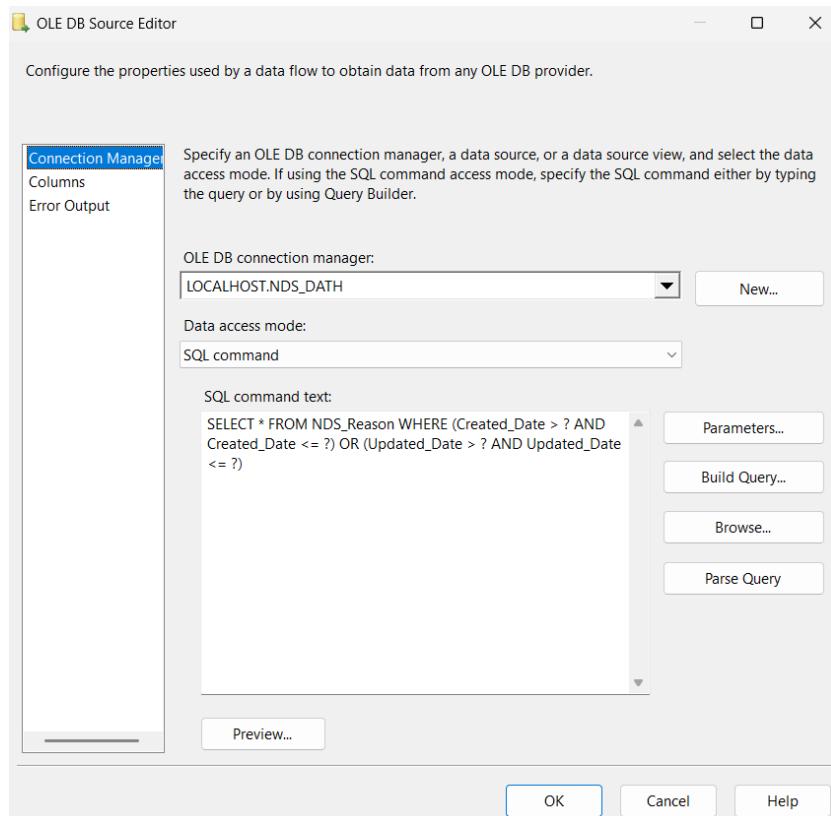
Bước 2.4: Khi Airline\_Name thay đổi:

- Cập nhật bản ghi cũ trong Dim\_Airport với Is\_Active = 0.
- Chèn bản ghi mới với khóa thay thế tự tăng và Is\_Active = 1.

### **c. Dim\_Reason**



**Bước 1:** Sử dụng OLE DB Source kết nối đến bảng NDS\_Reason. Truy vấn các bản ghi có CreatedDate hoặc UpdatedDate nằm trong khoảng LSET -> CET. Dữ liệu đầu ra gồm các lý do mới hoặc đã thay đổi kể từ lần ETL trước.



## Bước 2: Xử lý thay đổi bằng Slowly Changing Dimension (SCD)

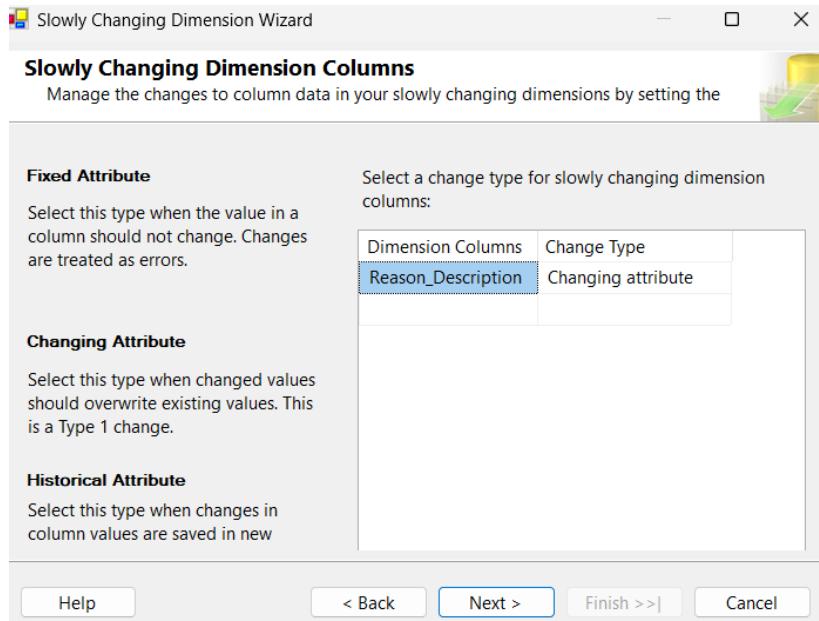
Bước 2.1: Dựa dữ liệu từ NDS\_Reason vào Slowly Changing Dimension Component.

Bước 2.2: Xác định Business Key là Reason\_Code.

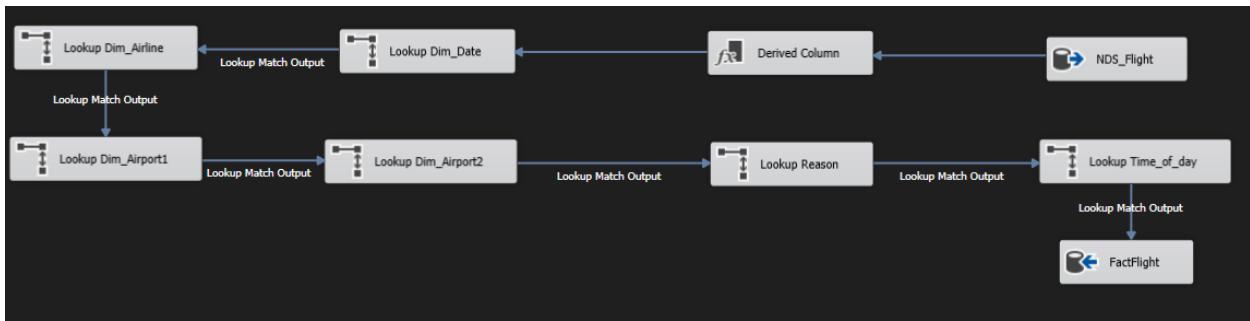
Bước 2.3: Cấu hình thuộc tính Reason\_Description là Changing Attribute (SCD Type 1).

Bước 2.4: Khi Reason\_Description thay đổi:

- Thực hiện UPDATE trực tiếp bản ghi hiện có trong Dim\_Reason.
- Giá trị cũ của Reason\_Description bị ghi đè bởi giá trị mới nhất.



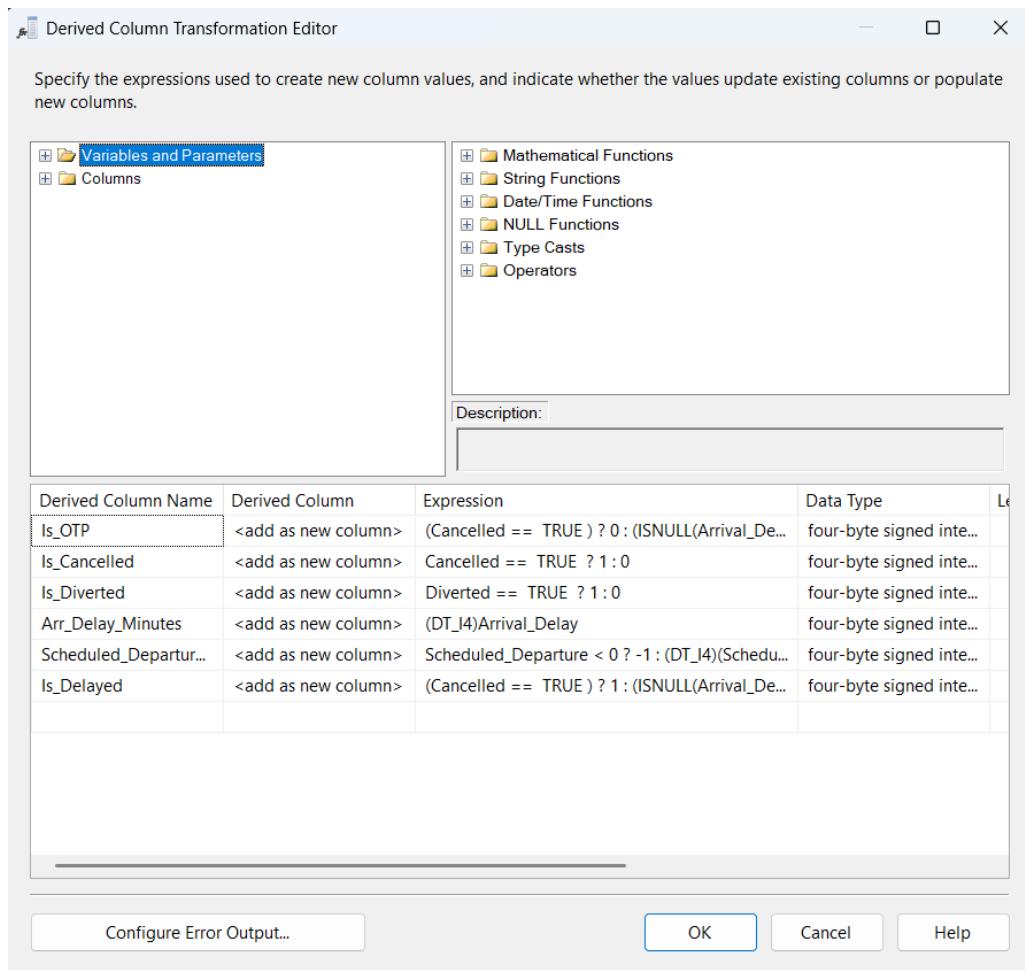
## 2.2.2. Nạp bảng sự kiện (Fact)



**Bước 1:** Lọc dữ liệu tăng trưởng theo khoảng thời gian LSET → CET dựa trên Created\_Date và Updated\_Date. Trích xuất các nhóm dữ liệu chính:

- Business Key: Iata\_Airline, Origin\_Airport, Destination\_Airport, Cancellation\_Reason
- Thông tin thời gian: Năm, Tháng, Ngày, Scheduled\_Departure
- Measure và Delay Breakdown
- Flag: Cancelled, Diverted

**Bước 2:** Dù Thực hiện các công việc như chuẩn hóa dữ liệu, tạo các cột mới, sửa định dạng, tính toán bằng Derived Column

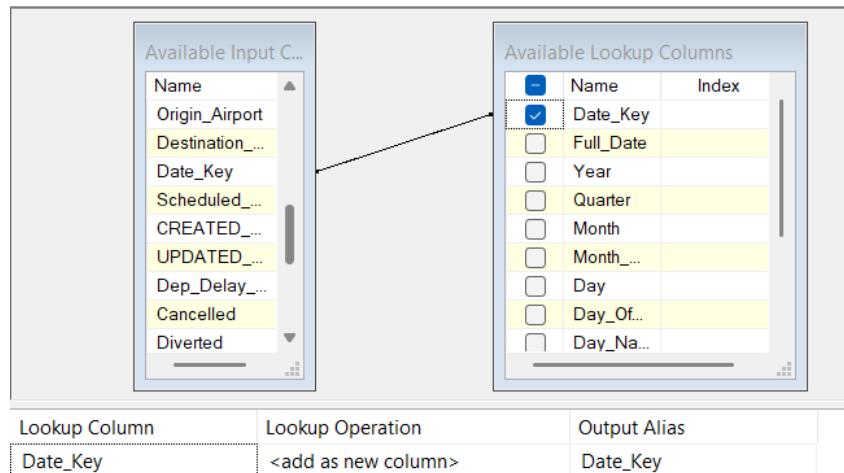


Thuộc tính	Chuẩn hóa / Xử lý trong Derived Column
<b>Is OTP</b>	$(\text{Cancelled} == \text{TRUE}) ? 0 : (\text{ISNULL}(\text{Arrival\_Delay}) ? 0 : ((\text{DT\_I4})\text{Arrival\_Delay} < 15 ? 1 : 0))$ <ul style="list-style-type: none"> <li>- Nếu chuyến bay bị hủy <math>\Rightarrow</math> Is OTP = 0</li> <li>- Nếu không hủy: <math>\text{Arrival\_Delay} &lt; 15 \Rightarrow 1</math>, ngược lại <math>\Rightarrow 0</math></li> </ul>
<b>Is_Cancelled</b>	$\text{Cancelled} == \text{TRUE} ? 1 : 0$ <p>Chuẩn hóa trạng thái hủy chuyến về dạng số (0/1)</p>
<b>Is_Diverted</b>	$\text{Diverted} == \text{TRUE} ? 1 : 0$ <p>Chuẩn hóa trạng thái chuyến hướng về dạng số (0/1)</p>
<b>Arr_Delay_Minutes</b>	$(\text{DT\_I4})\text{Arrival\_Delay}$ <p>Ép kiểu thời gian trẽ đến về số nguyên 4 byte</p>

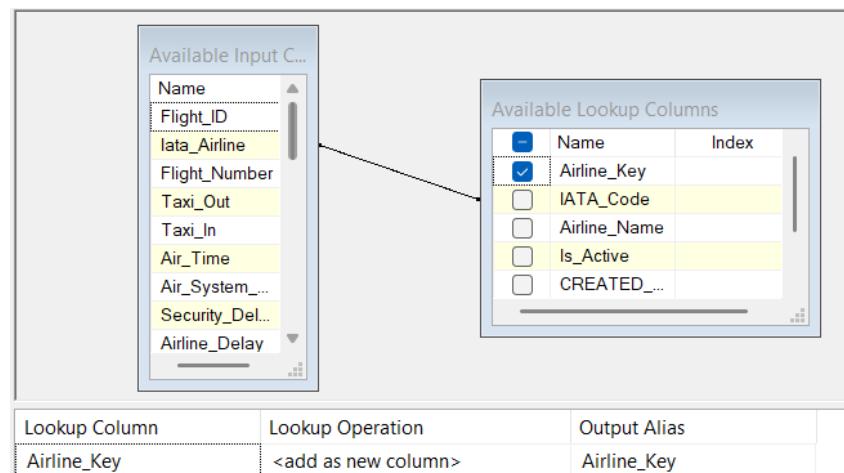
Scheduled_Departure_Hour	<p>Scheduled_Departure &lt; 0 ? -1 :  <math>(DT_I4)(Scheduled_Departure / 100)</math></p> <ul style="list-style-type: none"> <li>Nếu giá trị không hợp lệ =&gt; gán -1</li> <li>Ngược lại, tách giờ từ định dạng HHmm</li> </ul>
Is_Delayed	<p>(Cancelled == TRUE) ? 1 : (ISNULL(Arrival_Delay)  ? 0 : ((DT_I4)Arrival_Delay &gt;= 15 ? 1 : 0))</p> <ul style="list-style-type: none"> <li>Nếu chuyến bay bị hủy =&gt; Is_Delayed = 1</li> <li>Nếu không hủy: Arrival_Delay &gt;= 15 =&gt; 1,  ngược lại =&gt; 0</li> </ul>

**Bước 3:** Sau khi chuẩn hóa, mỗi bản ghi chuyến bay **bắt đầu lượt kiểm tra sự tồn tại trong các bảng Dimension** thông qua Lookup, cụ thể:

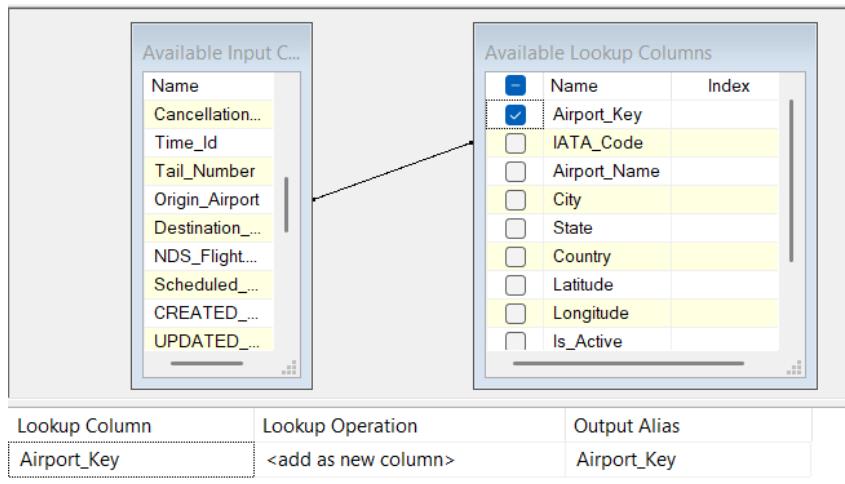
Bước 3.1: Kiểm tra Date\_Key có tồn tại trong Dim\_Date để lấy Date\_Key.



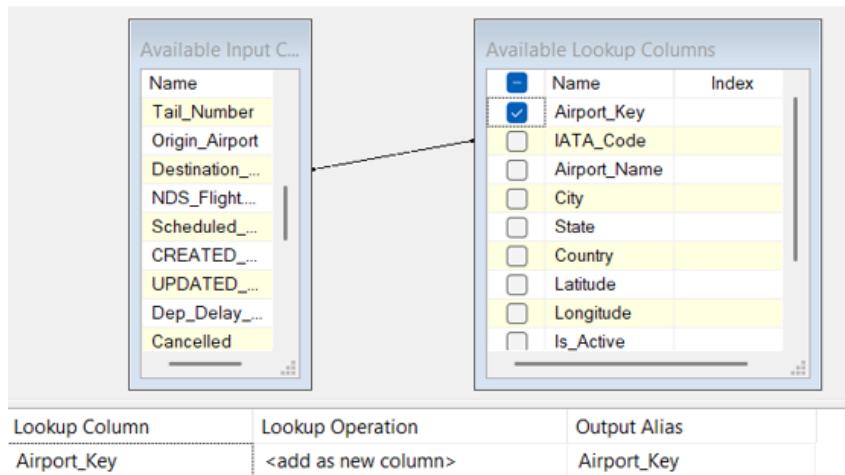
Bước 3.2: Kiểm tra mã hãng bay (Iata\_Airline) có tồn tại trong Dim\_Airline.IATA\_Code để lấy Airline\_Key.



Bước 3.3: Kiểm tra sân bay xuất phát (Origin\_Airport) có tồn tại trong Dim\_Airport.IATA\_Code để lấy Origin\_Airport\_Key.



Bước 3.4: Kiểm tra sân bay đến (Destination\_Airport) có tồn tại trong Dim\_Airport.IATA\_Code (Is\_Active = 1) để lấy Dest\_Airport\_Key;



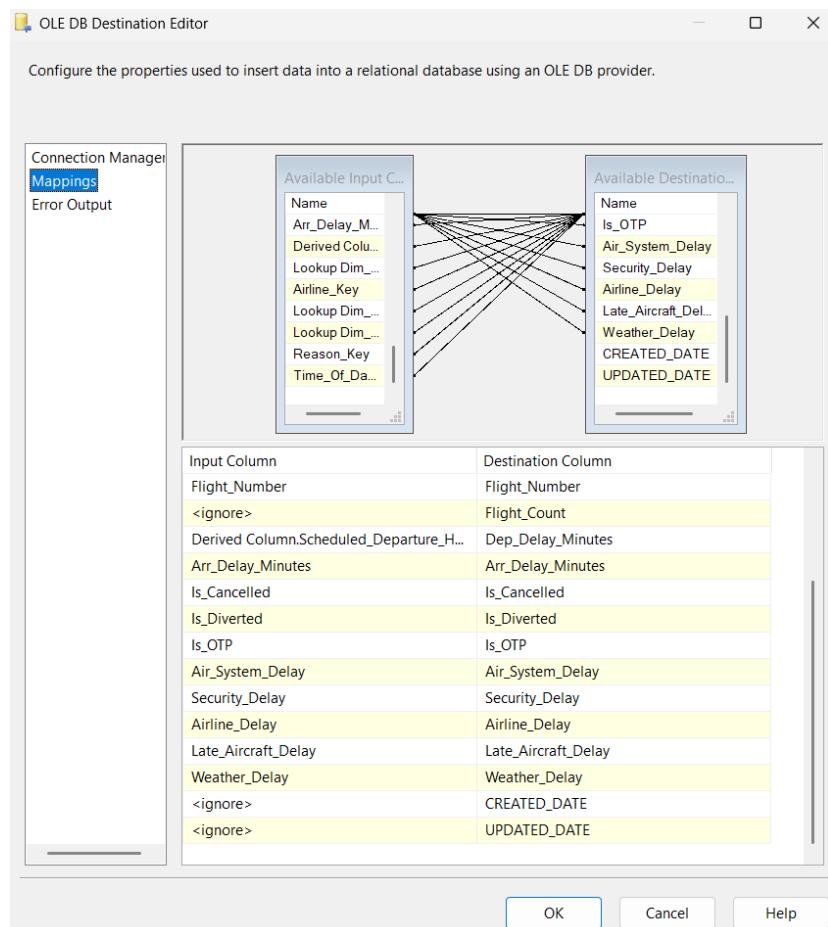
Bước 3.5: Nếu chuyến bay bị hủy, kiểm tra Cancellation\_Reason có tồn tại trong Dim\_Reason.Reason\_Code để lấy Reason\_Key;

Lookup Column	Lookup Operation	Output Alias
Reason_Key	<add as new column>	Reason_Key

Bước 3.6: Kiểm tra giờ khởi hành theo khung giờ (Scheduled\_Departure\_Hour) có tồn tại trong Dim\_TimeOfDay.Hour để lấy TimeOfDay\_Key.

Lookup Column	Lookup Operation	Output Alias
TimeOfDay_Key	<add as new column>	TimeOfDay_Key

**Bước 4:** Surrogate Key lấy được từ Lookup cùng với các measure, KPI và cột quản trị sẽ được map trực tiếp vào bảng Fact\_Flight theo sơ đồ ánh xạ Fact như bên dưới và ghi vào DDS bằng OLE DB Destination.

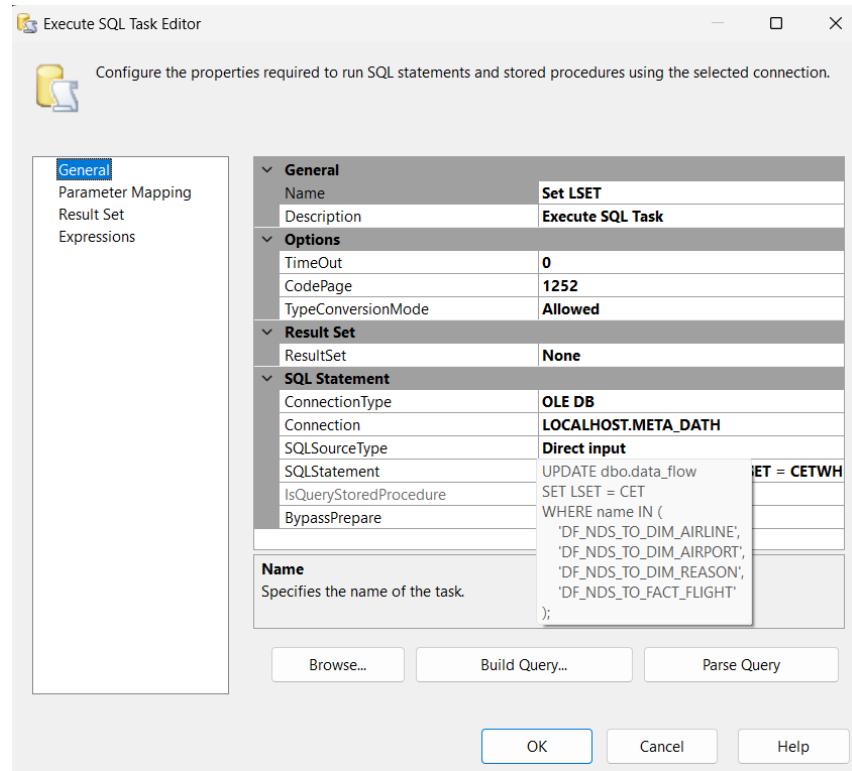


Input Column (Data Flow)	Destination Column (Fact_Flight)
Lookup_Dim_Date.Date_Key	Date_Key
Lookup_Dim_Airline.Airline_Key	Airline_Key
Lookup_Dim_Airport1.Airport_Key	Origin_Airport_Key
Lookup_Dim_Airport2.Airport_Key	Dest_Airport_Key
Reason_Key	Reason_Key
Time_Of_Day_Key	Time_Of_Day_Key
Flight_Number	Flight_Number
(constant / ignore)	Flight_Count
Arr_Delay_Minutes	Arr_Delay_Minutes
Dep_Delay_Minutes	Dep_Delay_Minutes

Is_Cancelled	Is_Cancelled
Is_Diverted	Is_Diverted
Is OTP	Is OTP
Air_System_Delay	Air_System_Delay
Security_Delay	Security_Delay
Airline_Delay	Airline_Delay
Late_Aircraft_Delay	Late_Aircraft_Delay
Weather_Delay	Weather_Delay
CREATED_DATE	CREATED_DATE
UPDATED_DATE	UPDATED_DATE

### **Giai đoạn 3: Hoàn tất (Set LSET)**

Sau khi toàn bộ các Data Flow (NDS -> Dimension và NDS -> Fact) chạy thành công, hệ thống thực hiện cập nhật Last Successful Extract Time (LSET) bằng Current Extract Time (CET) trong bảng metadata dbo.data\_flow.



### 2.3. Kết quả

- Dim\_Airline

57  SELECT \* FROM DIM\_AIRLINE  
 58  SELECT \* FROM DIM\_AIRPORT

122 %

	Airline_Key	IATA_Code	Airline_Name	Is_Active	CREATED_DATE	UPDATED_DATE
1	1	AA	American Airlines Inc.	1	2025-12-16 18:31:03.083	2025-12-16 18:31:03.083
2	2	AS	Alaska Airlines Inc.	1	2025-12-16 18:31:03.083	2025-12-16 18:31:03.083
3	3	B6	JetBlue Airways	1	2025-12-16 18:31:03.087	2025-12-16 18:31:03.087
4	4	DL	Delta Air Lines Inc.	1	2025-12-16 18:31:03.087	2025-12-16 18:31:03.087
5	5	EV	Atlantic Southeast Airlines	1	2025-12-16 18:31:03.087	2025-12-16 18:31:03.087
6	6	F9	Frontier Airlines Inc.	1	2025-12-16 18:31:03.087	2025-12-16 18:31:03.087
7	7	HA	Hawaiian Airlines Inc.	1	2025-12-16 18:31:03.087	2025-12-16 18:31:03.087
8	8	MQ	American Eagle Airlines Inc.	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
9	9	NK	Spirit Air Lines	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
10	10	OO	Skywest Airlines Inc.	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
11	11	UA	United Air Lines Inc.	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
12	12	US	US Airways Inc.	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
13	13	VX	Virgin America	1	2025-12-16 18:31:03.090	2025-12-16 18:31:03.090
14	14	WN	Southwest Airlines Co.	1	2025-12-16 18:31:03.093	2025-12-16 18:31:03.093

- Dim\_Airport

58 | `SELECT * FROM DIM_AIRPORT`

59 | `SELECT * FROM DIM_DATE`

60 | `SELECT * FROM DIM_REASON`

61 | `SELECT * FROM DIM_TIME_OF_DAY`

122 %

Results Messages

	Airport_Key	IATA_Code	Airport_Name	City	State	Country	Latitude	Longitude	Is_Active	CREATED_DATE	UPDATED_DATE
1	1	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.6523590087891	-75.4403991699219	1	2025-12-16 18:31:03.453	2025-12-16 18:31:03.453
2	2	ABI	Abilene Regional Airport	Abilene	TX	USA	32.411319732666	-99.6819000244141	1	2025-12-16 18:31:03.457	2025-12-16 18:31:03.457
3	3	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.0402183532715	-106.609191894531	1	2025-12-16 18:31:03.457	2025-12-16 18:31:03.457
4	4	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.4490585327148	-98.4218292236328	1	2025-12-16 18:31:03.457	2025-12-16 18:31:03.457
5	5	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552050535889	-84.1944732666016	1	2025-12-16 18:31:03.457	2025-12-16 18:31:03.457
6	6	ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.2530517578125	-70.0601806640625	1	2025-12-16 18:31:03.457	2025-12-16 18:31:03.457
7	7	ACT	Waco Regional Airport	Waco	TX	USA	31.6112099780273	-97.2305221557617	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
8	8	ACV	Arcata Airport	Arcata/Eureka	CA	USA	40.9781188964844	-124.108619689941	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
9	9	ACY	Atlantic City International Airport	Atlantic City	NJ	USA	39.4575005640603	-74.5771713256836	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
10	10	ADK	Adak Airport	Adak	AK	USA	51.8779602050781	-176.646026611328	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
11	11	ADQ	Kodiak Airport	Kodiak	AK	USA	57.499694824219	-152.493865966797	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
12	12	AEX	Alexandria International Airport	Alexandria	LA	USA	31.3273696899414	-92.5485610961914	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
13	13	AGS	Augusta Regional Airport (Bush Field)	Augusta	GA	USA	33.3699607849121	-81.9645004272461	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
14	14	AKN	King Salmon Airport	King Salmon	AK	USA	58.6767907741699	-156.649215698242	1	2025-12-16 18:31:03.460	2025-12-16 18:31:03.460
15	15	ALB	Albany International Airport	Albany	NY	USA	42.748119354248	-73.802978515628	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
16	16	ALO	Watertown Regional Airport	Watertown	IA	USA	42.5570793151855	-92.4003372192383	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
17	17	AMA	Rick Husband Amarillo International Airport	Amarillo	TX	USA	35.2193717956543	-101.705932617188	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
18	18	ANC	Ted Stevens Anchorage International Airport	Anchorage	AK	USA	61.1743202209473	-149.996185302734	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
19	19	APN	Alpena County Regional Airport	Alpena	MI	USA	45.0780715942383	-83.5602874755859	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
20	20	ASE	Aspen-Pitkin County Airport	Aspen	CO	USA	39.2231597900391	-106.86885070808	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
21	21	ATL	Hartsfield-Jackson Atlanta International Airport	Atlanta	GA	USA	33.6404418945313	-84.4269409179688	1	2025-12-16 18:31:03.463	2025-12-16 18:31:03.463
22	22	ATW	Appleton International Airport	Appleton	WI	USA	44.2574081420898	-88.5194778442383	1	2025-12-16 18:31:03.467	2025-12-16 18:31:03.467
23	23	AUS	Austin-Bergstrom International Airport	Austin	TX	USA	30.1945304870605	-97.6698684692383	1	2025-12-16 18:31:03.467	2025-12-16 18:31:03.467

## - Dim\_Date

59 | `SELECT * FROM DIM_DATE`

60 | `SELECT * FROM DIM_REASON`

61 | `SELECT * FROM DIM_TIME_OF_DAY`

122 %

Results Messages

	Date_Key	Full_Date	Year	Quarter	Month	Month_Name	Day	Day_Of_Week	Day_Name	Is_Weekend	Season	CREATED_DATE	UPDATED_DATE
1	20050101	2005-01-01	2005	1	1	January	1	7	Saturday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
2	20050102	2005-01-02	2005	1	1	January	2	1	Sunday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
3	20050103	2005-01-03	2005	1	1	January	3	2	Monday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
4	20050104	2005-01-04	2005	1	1	January	4	3	Tuesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
5	20050105	2005-01-05	2005	1	1	January	5	4	Wednesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
6	20050106	2005-01-06	2005	1	1	January	6	5	Thursday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
7	20050107	2005-01-07	2005	1	1	January	7	6	Friday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
8	20050108	2005-01-08	2005	1	1	January	8	7	Saturday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
9	20050109	2005-01-09	2005	1	1	January	9	1	Sunday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
10	20050110	2005-01-10	2005	1	1	January	10	2	Monday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
11	20050111	2005-01-11	2005	1	1	January	11	3	Tuesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
12	20050112	2005-01-12	2005	1	1	January	12	4	Wednesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
13	20050113	2005-01-13	2005	1	1	January	13	5	Thursday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
14	20050114	2005-01-14	2005	1	1	January	14	6	Friday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
15	20050115	2005-01-15	2005	1	1	January	15	7	Saturday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
16	20050116	2005-01-16	2005	1	1	January	16	1	Sunday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
17	20050117	2005-01-17	2005	1	1	January	17	2	Monday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
18	20050118	2005-01-18	2005	1	1	January	18	3	Tuesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
19	20050119	2005-01-19	2005	1	1	January	19	4	Wednesday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
20	20050120	2005-01-20	2005	1	1	January	20	5	Thursday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
21	20050121	2005-01-21	2005	1	1	January	21	6	Friday	0	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
22	20050122	2005-01-22	2005	1	1	January	22	7	Saturday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433
23	20050123	2005-01-23	2005	1	1	January	23	1	Sunday	1	Winter	2025-12-16 18:30:47.433	2025-12-16 18:30:47.433

## - Dim\_Reason

```

60 | SELECT * FROM DIM_REASON
61 | SELECT * FROM DIM_TIME_OF_DAY
122 % ▾
Results Messages

```

	Reason_Key	Reason_Code	Reason_Description	CREATED_DATE	UPDATED_DATE
1	1	N	Not Cancelled	2025-12-16 18:30:47.813	2025-12-16 18:30:47.813
2	2	A	Airline/Carrier	2025-12-16 18:31:03.717	2025-12-16 18:31:03.717
3	3	B	Weather	2025-12-16 18:31:03.720	2025-12-16 18:31:03.720
4	4	C	National Air System	2025-12-16 18:31:03.720	2025-12-16 18:31:03.720

#### - Dim\_TimeOfDay

```

61 | SELECT * FROM DIM_TIME_OF_DAY
62 |
122 % ▾
Results Messages

```

	TimeOfDay_Key	Hour	TimeOfDay_Name	CREATED_DATE	UPDATED_DATE
1	1	0	Night	2025-12-16 18:30:47.797	2025-12-16 18:30:47.797
2	2	1	Night	2025-12-16 18:30:47.797	2025-12-16 18:30:47.797
3	3	2	Night	2025-12-16 18:30:47.800	2025-12-16 18:30:47.800
4	4	3	Night	2025-12-16 18:30:47.800	2025-12-16 18:30:47.800
5	5	4	Night	2025-12-16 18:30:47.800	2025-12-16 18:30:47.800
6	6	5	Early Morning	2025-12-16 18:30:47.800	2025-12-16 18:30:47.800
7	7	6	Early Morning	2025-12-16 18:30:47.800	2025-12-16 18:30:47.800
8	8	7	Early Morning	2025-12-16 18:30:47.803	2025-12-16 18:30:47.803
9	9	8	Early Morning	2025-12-16 18:30:47.803	2025-12-16 18:30:47.803
10	10	9	Morning	2025-12-16 18:30:47.803	2025-12-16 18:30:47.803
11	11	10	Morning	2025-12-16 18:30:47.803	2025-12-16 18:30:47.803
12	12	11	Morning	2025-12-16 18:30:47.807	2025-12-16 18:30:47.807
13	13	12	Afternoon	2025-12-16 18:30:47.807	2025-12-16 18:30:47.807
14	14	13	Afternoon	2025-12-16 18:30:47.807	2025-12-16 18:30:47.807
15	15	14	Afternoon	2025-12-16 18:30:47.807	2025-12-16 18:30:47.807
16	16	15	Afternoon	2025-12-16 18:30:47.807	2025-12-16 18:30:47.807
17	17	16	Afternoon	2025-12-16 18:30:47.810	2025-12-16 18:30:47.810
18	18	17	Evening	2025-12-16 18:30:47.810	2025-12-16 18:30:47.810
19	19	18	Evening	2025-12-16 18:30:47.810	2025-12-16 18:30:47.810

## VII. Tự động hóa ETL

### 1. Giải pháp

Nhóm sử dụng SQL Server Agent Job để tự động hóa việc thực thi package ETL SSIS. Công cụ này cho phép cấu hình lịch chạy định kỳ, theo dõi trạng thái thực thi và ghi nhận lỗi trong quá trình xử lý.

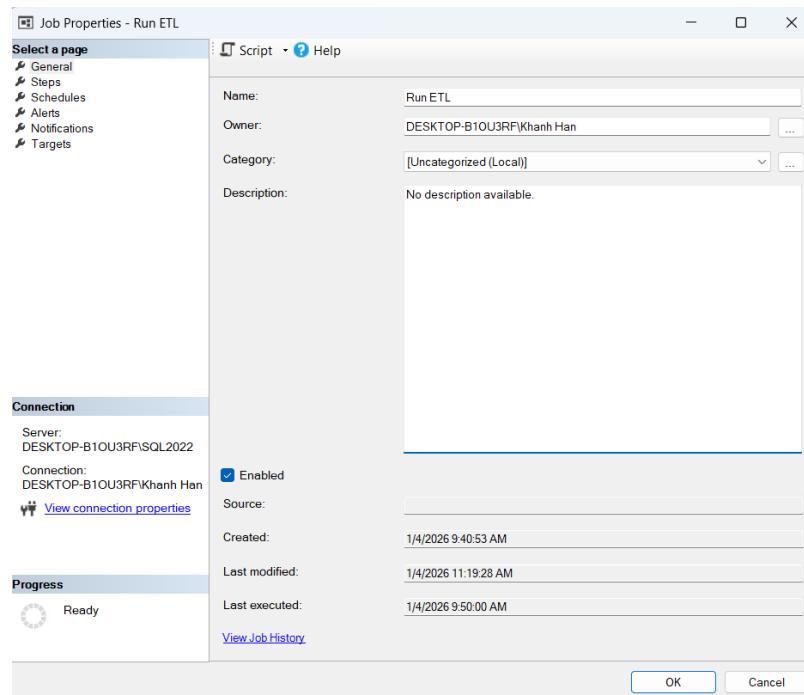
SQL Server Agent Job cho phép:

- Tự động kích hoạt các SSIS Package
- Cấu hình lịch chạy định kỳ (theo ngày, tuần hoặc theo nhu cầu)
- Theo dõi trạng thái thực thi và ghi nhận lịch sử chạy (Job History)

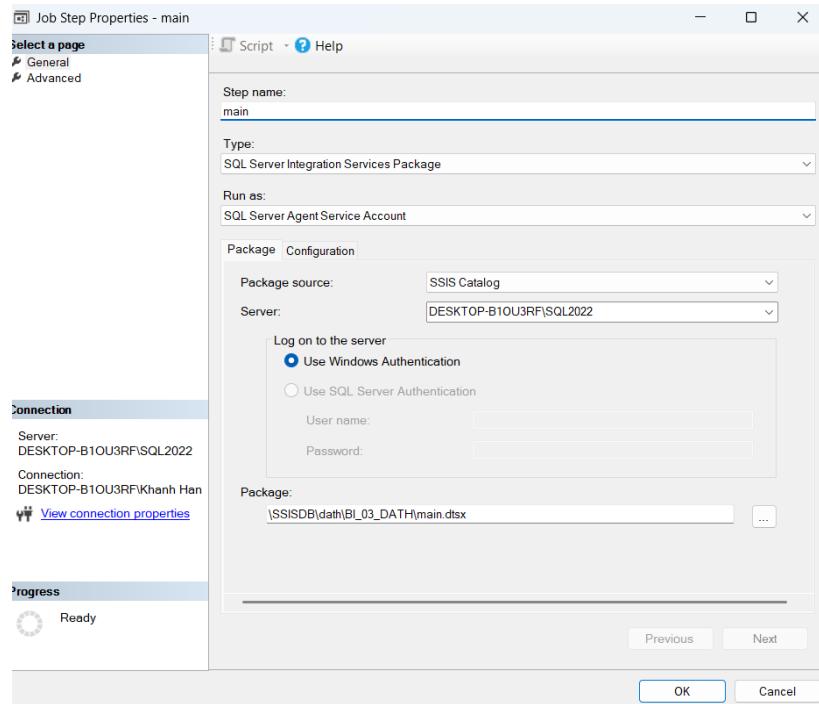
Giải pháp này phù hợp với hệ thống ETL được xây dựng bằng SQL Server Integration Services (SSIS) trong đồ án.

## 2. Quy trình

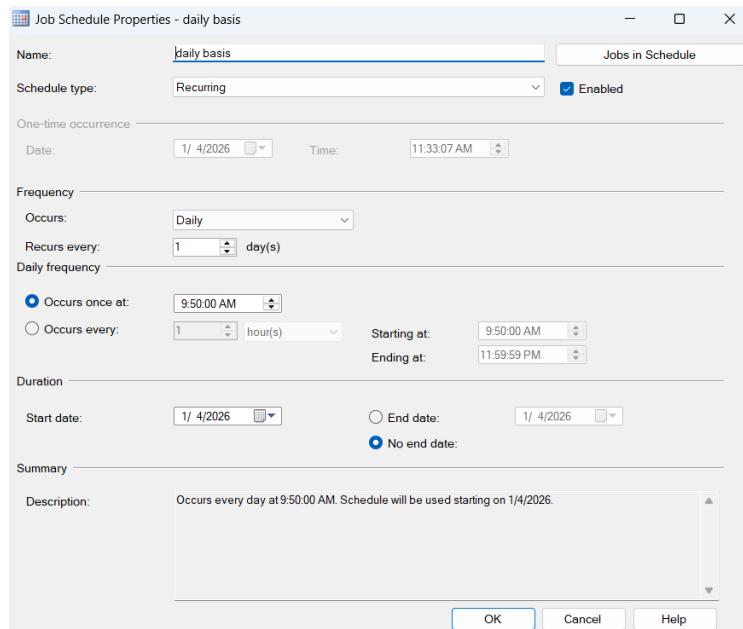
- Tạo một Job mới trong SQL Server Agent



- Cấu hình Job Step với loại tác vụ là SQL Server Integration Services Package



- Cấu hình Schedule để job chạy tự động theo lịch



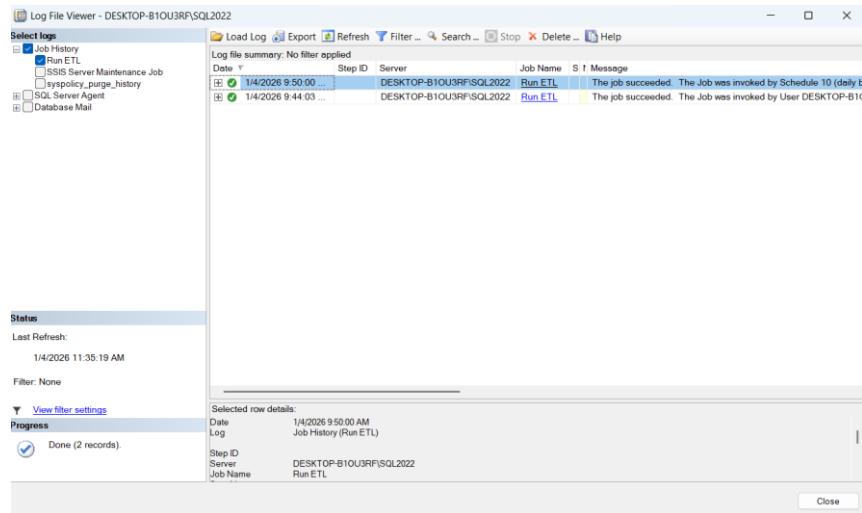
### 3. Kết quả

SQL Server Agent cung cấp cơ chế giám sát thông qua Job History, cho phép theo dõi chi tiết trạng thái của từng lần thực thi ETL, bao gồm:

- Thời gian bắt đầu và kết thúc

- Trạng thái thành công (Success) hoặc thất bại (Failed)
- Thông báo lỗi chi tiết trong trường hợp xảy ra sự cố

Ngoài ra, SSIS cũng ghi nhận log thực thi, hỗ trợ việc phân tích và khắc phục lỗi khi cần thiết.



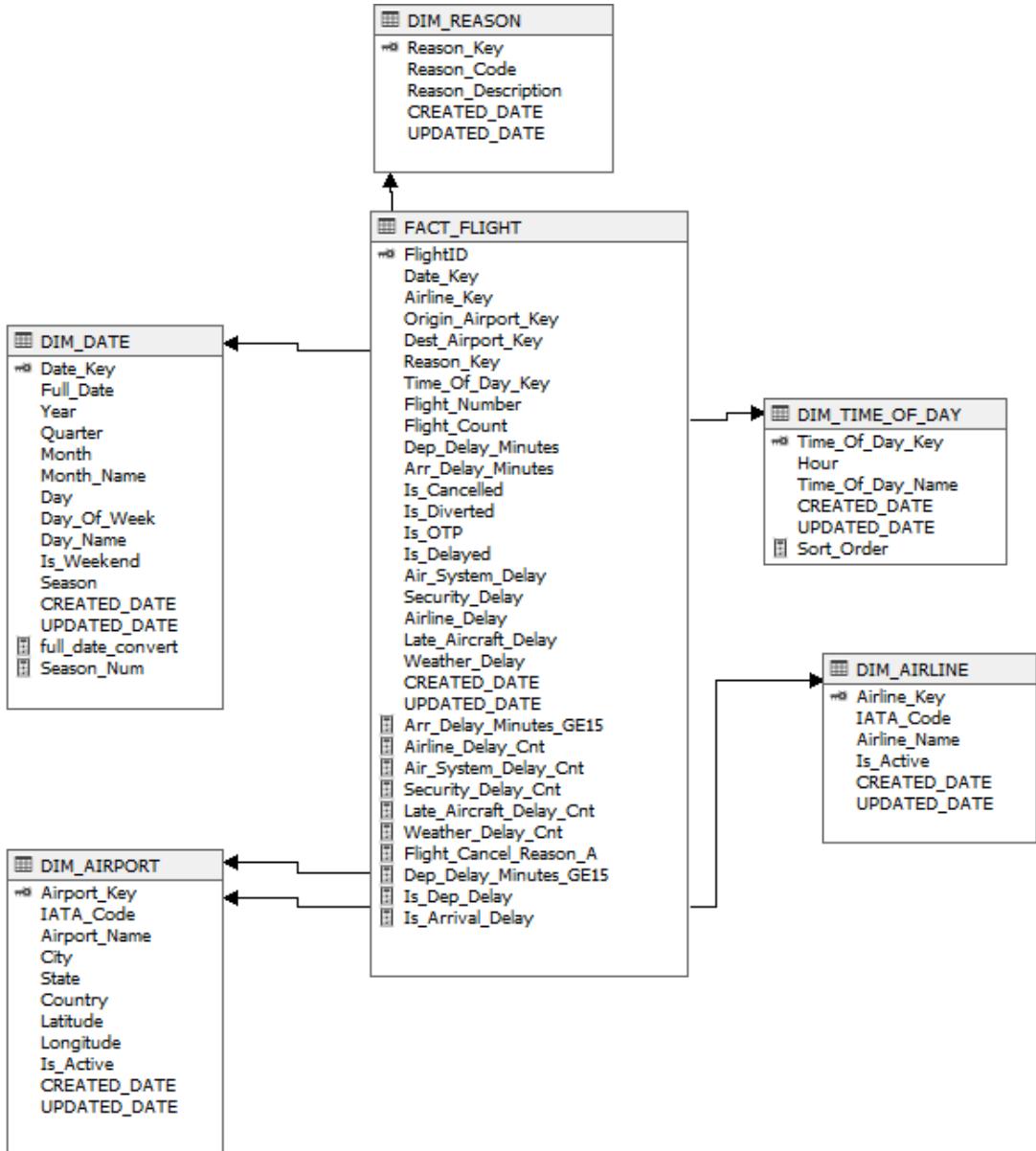
## VIII. Olap Cube

### 1. Mục tiêu

Mục tiêu của bước thiết kế OLAP Cube là xây dựng mô hình phân tích đa chiều từ dữ liệu đã được làm sạch và chuẩn hóa trong DDS, phục vụ cho việc phân tích, tổng hợp và truy vấn dữ liệu theo nhiều chiều khác nhau với hiệu năng cao.

### 2. Data source view (DSV)

Data Source View (DSV) được sử dụng để trích xuất các bảng cần thiết từ cơ sở dữ liệu DDS, làm nền tảng cho việc xây dựng các dimension và measure group trong OLAP Cube.



- Các bảng được sử dụng trong data source view:

Tên bảng	Mô tả
DIM_DATE	Phân tích dữ liệu theo thời gian
DIM_TIME_OF_DAY	Phân tích theo khung giờ trong ngày
DIM_AIRLINE	Phân tích theo hãng hàng không
DIM_AIRPORT	Phân tích theo sân bay
DIM_REASON	Phân tích theo lý hủy chuyến

FACT_FLIGHT	Lưu trữ dữ liệu chuyến bay, là cơ sở tạo Measure Group trong OLAP Cube
-------------	--

- Bảng DIM\_AIRPORT phục vụ hai vai trò khác nhau:
  - o Sân bay khởi hành (Origin Airport)
  - o Sân bay đến (Destination Airport)

### 3. OLAP cube

OLAP Cube DDS DATH được xây dựng trên nền Data Source View DDS DATH.dsv, sử dụng SQL Server Analysis Services hỗ trợ phân tích dữ liệu chuyến bay theo mô hình đa chiều.

Cube bao gồm một measure group chính: FACT\_FLIGHT (là bảng FACT\_FLIGHT trong DDS)

- Các measure được xây dựng từ bảng FACT\_FLIGHT:

Measure	Ý nghĩa
Flight Count	Tổng số chuyến bay (measure đếm, rất quan trọng để làm mẫu số khi tính tỷ lệ delay/hủy).
Dep Delay Minutes	Tổng số phút trễ khi khởi hành (Departure Delay)
Arr Delay Minutes	Tổng số phút trễ khi đến nơi (Arrival Delay)
Is Cancelled	Measure nhị phân (0/1) cho biết chuyến bay có bị hủy hay không; thường dùng để đếm số chuyến bị hủy.
Is Diverted	Measure nhị phân (0/1) cho biết chuyến bay có bị chuyển hướng (diverted) hay không.
Is OTP	On-Time Performance: chuyến bay đúng giờ (arrival delay < 15 phút).
Is Delayed	Measure nhị phân (0/1) cho biết chuyến bay có bị trễ hay không. (Tính cả những chuyến bị cancel).
Air System Delay	Tổng số phút trễ do hệ thống điều hành không lưu

Security Delay	Tổng số phút trễ do nguyên nhân an ninh (security).
Airline Delay	Tổng số phút trễ do hãng hàng không (crew, bảo trì, vận hành...)
Late Aircraft Delay	Tổng số phút trễ do máy bay đến muộn từ chuyến trước
Weather Delay	Tổng số phút trễ do thời tiết
Arr Delay Minutes GE15	Tổng số phút trễ khi đến nơi, chỉ tính các chuyến trễ $\geq 15$ phút
Weather Delay Cnt	Số chuyến bay bị trễ do thời tiết
Late Aircraft Delay Cnt	Số chuyến bay bị trễ do máy bay đến muộn
Security Delay Cnt	Số chuyến bay bị trễ do an ninh
Air System Delay Cnt	Số chuyến bay bị trễ do hệ thống không lưu
Airline Delay Cnt	Số chuyến bay bị trễ do hãng hàng không
Flight Cancel Reason A	Số chuyến bay bị hủy do lý do A (Carrier – hãng hàng không)
Dep Delay Minutes GE15	Tổng số phút trễ khởi hành, chỉ tính các chuyến trễ $\geq 15$ phút
Is Dep Delay	Measure nhị phân (0/1) cho biết chuyến bay có bị trễ khi khởi hành hay không. (Chỉ tính các chuyến thật sự trễ, không tính các chuyến cancel)
Is Arrival Delay	Measure nhị phân (0/1) cho biết chuyến bay có bị trễ khi đến nơi hay không. (Chỉ tính các chuyến thật sự trễ, không tính các chuyến cancel)

- Các calculated measure được xây dựng:

Measure không lưu trực tiếp trong fact table, mà được tính toán động bằng MDX dựa trên các measure khác và ngữ cảnh (dimension, hierarchy, filter).

Measure	Công thức MDX	Ý nghĩa
---------	---------------	---------

Percent Cancel Reason A	[Measures].[Flight Cancel Reason A]/[Measures].[Is Cancelled]	Tỷ lệ chuyến bay bị hủy do lý do A (Carrier). Dùng để phân tích nguyên nhân hủy.
Percent OTP	[Measures].[Is OTP] / [Measures].[Flight Count]	Tỷ lệ chuyến bay đúng giờ (On-Time Performance)
Percent Arr Delay	[Measures].[Is Delayed] / [Measures].[Flight Count]	Tỷ lệ chuyến bay bị trễ khi đến nơi.
Percent Cancel	[Measures].[Is Cancelled] / [Measures].[Flight Count]	Tỷ lệ chuyến bay bị hủy
Percent Dep Delay	[Measures].[Is Dep Delay]/ [Measures].[Flight Count]	Tỷ lệ chuyến bay bị trễ khi khởi hành
Avg Arrival Delay	[Measures].[Arr Delay Minutes GE15] /[Measures].[Is Arrival Delay]	Thời gian trễ đến trung bình ( $\geq 15$ phút)
Avg Departure Delay	[Measures].[Dep Delay Minutes GE15] / [Measures].[Is Dep Delay]	Thời gian trễ khởi hành trung bình ( $\geq 15$ phút)
Avg Air System Delay	[Measures].[Air System Delay] / [Measures].[Air System Delay Cnt]	Số phút trễ trung bình do hệ thống không lưu
Avg Security Delay	[Measures].[Security Delay] / [Measures].[Security Delay Cnt]	Số phút trễ trung bình do an ninh
Avg Airline Delay	[Measures].[Airline Delay]/[Measures].[Airline Delay Cnt]	Số phút trễ trung bình do hãng hàng không
Avg Late Aircraft Delay	[Measures].[Late Aircraft Delay] /[Measures].[Late Aircraft Delay Cnt]	Số phút trễ trung bình do máy bay đến muộn
Avg Weather Delay	[Measures].[Weather Delay] /	Số phút trễ trung bình do thời tiết

	[Measures].[Weather Delay Cnt]	
--	--------------------------------	--

- Cube sử dụng các dimension sau:

Dimension	Vai trò
DIM DATE	Phân tích theo thời gian
DIM TIME OF DAY	Phân tích theo khung giờ
DIM AIRLINE	Phân tích theo hãng hàng không
DIM REASON	Phân tích theo nguyên nhân trễ/hủy
Origin Airport	Phân tích theo sân bay khởi hành
Dest Airport	Phân tích theo sân bay đến

- Kiểm tra và khai thác OLAP Cube:

The screenshot shows the DDS DATH interface. On the left, there's a sidebar with 'DDS DATH' at the top, followed by 'Metadata', 'Search Model', 'Measure Group: <All>', and a large list of measures including 'Airline Delay', 'Arr Delay Minutes', 'Dep Delay Minutes', 'Flight Count', 'Is Arrival Delay', 'Is Cancelled', 'Is Delayed', 'Is Dep Delay', 'Is Diverted', 'Is OTP', and 'Late Aircraft Delay'. Below this is a section for 'Calculated Members'. On the right, there's a main area with tabs for 'Dimension', 'Hierarchy', 'Operator', and 'Filter Expression'. Under the 'Dimension' tab, it says '<Select dimension>'. Below these tabs is a table with data from the cube.

Airline Key	Flight Count	Is Cancelled	Is Delayed	Is Diverted	Is OTP
American Airlines Inc.	12423	191	2419	38	10004
Alaska Airlines Inc.	3033	9	397	7	2636
JetBlue Airways	4618	75	1081	13	3537
Delta Air Lines Inc.	15051	75	2084	32	12967
Atlantic Southeast Airlines	9843	268	2233	27	7610
Frontier Airlines Inc.	1528	16	405	4	1123
Hawaiian Airlines Inc.	1303	1	166	2	1137
American Eagle Airlines Inc.	5058	244	1281	12	3777
Spirit Air Lines	2013	40	623	1	1390
Skywest Airlines Inc.	10155	169	2017	30	8138
United Air Lines Inc.	8918	103	1938	24	6980
US Airways Inc.	3397	87	693	10	2704
Virgin America	1058	7	209	1	849
Southwest Airlines Co.	21602	273	4247	61	17355

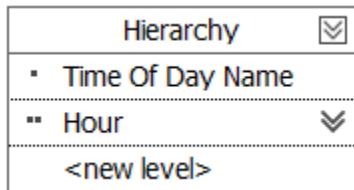
- Dimension DIM AIRLINE được sử dụng để phân tích dữ liệu theo hãng hàng không, kết hợp với các measure như Flight Count, Is Cancelled, Is Diverted và Is OTP.
- Kết quả cho thấy OLAP Cube trả về số liệu tổng hợp chính xác cho từng hãng hàng không, phản ánh đầy đủ các chỉ số về số lượng chuyến bay, tình trạng hủy chuyến, chuyến hướng và tỷ lệ chuyến bay đúng giờ.

## 4. Phân cấp chiều

### a. Dim\_TimeOfDay

Dimension TIME OF DAY phục vụ phân tích dữ liệu chuyến bay theo khung giờ trong ngày. Việc phân tích theo thời điểm trong ngày giúp đánh giá sự khác biệt về lưu lượng chuyến bay, tình trạng trễ và hiệu suất đúng giờ tại các khoảng thời gian khác nhau.

Phân cấp chiều đã được xây dựng nhằm hỗ trợ thao tác drill-down trong OLAP Cube: Time Of Day Name → Hour



Trong đó:

- Time Of Day Name đại diện cho nhóm khung giờ trong ngày
- Hour đại diện cho từng giờ cụ thể trong khung giờ đó

Phân cấp này cho phép:

- Phân tích tổng quan theo từng khung giờ trong ngày
- Drill-down xuống mức chi tiết theo từng giờ cụ thể

Row Labels	Flight Count	Is Cancelled	Is OTP
<b>Early Morning</b>			
5	2024	39	1831
6	7008	141	6273
7	6767	95	5967
8	6557	95	5641
<b>Morning</b>			
9	6033	69	5147
10	6407	74	5401
11	6154	89	5153
<b>Afternoon</b>			
	<b>30093</b>	<b>404</b>	<b>23878</b>
<b>Evening</b>			
	<b>28567</b>	<b>551</b>	<b>21313</b>
<b>Night</b>			
	<b>390</b>	<b>1</b>	<b>330</b>
<b>Grand Total</b>		<b>100000</b>	<b>1558 80934</b>

## b. Dim\_Date

Dimension DATE được sử dụng trong OLAP Cube phục vụ phân tích dữ liệu chuyến bay theo thời gian. Dimension cho phép đánh giá xu hướng, so sánh và tổng hợp dữ liệu theo các mức thời gian khác nhau như năm, quý, tháng và mùa.

Có 2 phân cấp chiều, đáp ứng các nhu cầu phân tích thời gian khác nhau:

- Hierarchy 1 – Phân tích theo thời gian chuẩn: Year → Quarter → Month → Day



- Phân cấp này hỗ trợ:

- Phân tích dữ liệu theo từng năm
- Drill-down từ năm xuống quý, tháng và ngày cụ thể

Row Labels	Flight Count	Is Cancelled	Is Delayed	Is Diverted	Is OTP
2015					
1					
January					
1	241	5	58	0	183
2	289	6	84	2	205
3	266	3	130	0	136
4	282	10	146	0	136
5	285	9	92	0	193
6	264	5	110	1	154
7	269	5	84	1	185
8	276	16	76	2	200

- Hierarchy 1 – Phân tích theo mùa: Season → Month → Day



- Phân cấp này hỗ trợ:

- Phân tích dữ liệu theo mùa trong năm
- Đánh giá ảnh hưởng của yếu tố mùa vụ đến tình trạng trễ hoặc hủy chuyến

Row Labels	Flight Count	Is Cancelled	Is Delayed	Is Diverted	Is OTP
Spring					
March					
1	262	30	128	6	134
2	283	7	103	0	180
3	276	22	118	4	158
4	280	21	94	1	186
5	287	41	107	2	180
6	287	10	101	0	186
7	246	4	32	0	214
8	277	3	58	0	219
9	291	1	42	1	249

Trong Dimension DATE, một số thuộc tính như Month và Quarter có thể lặp lại giá trị giữa các năm khác nhau (ví dụ: Tháng 1, Quý 1 xuất hiện ở nhiều năm). Điều này có thể dẫn đến xung đột khóa khi xây dựng hierarchy trong OLAP Cube. Để giải quyết vấn đề này, Dimension DATE được cấu hình theo nguyên tắc Composite Key, trong đó:

- Thuộc tính Month được định danh bởi cặp khóa (Year, Month)
- Thuộc tính Quarter được định danh bởi cặp khóa (Year, Quarter)
- Thuộc tính Day được định danh bởi các khóa (Year, Month, Day)

<table border="1"> <thead> <tr> <th colspan="2">KeyColumns</th><th>(Collection)</th></tr> </thead> <tbody> <tr> <td>▪ DIM_DATE.Quarter (Integer)</td><td>DIM_DATE.Quarter (Integer)</td><td>...</td></tr> <tr> <td>▪ DIM_DATE.Year (Integer)</td><td>DIM_DATE.Year (Integer)</td><td></td></tr> <tr> <td>MemberNamesUnique</td><td>False</td><td></td></tr> <tr> <td>MembersWithData</td><td>NonLeafDataVisible</td><td></td></tr> <tr> <td>MembersWithDataCaption</td><td>MembersWithDataCaption</td><td></td></tr> <tr> <td>Name</td><td>Quarter</td><td></td></tr> <tr> <td>▪ NameColumn</td><td>DIM_DATE.Quarter (WChar)</td><td></td></tr> </tbody> </table>		KeyColumns		(Collection)	▪ DIM_DATE.Quarter (Integer)	DIM_DATE.Quarter (Integer)	...	▪ DIM_DATE.Year (Integer)	DIM_DATE.Year (Integer)		MemberNamesUnique	False		MembersWithData	NonLeafDataVisible		MembersWithDataCaption	MembersWithDataCaption		Name	Quarter		▪ NameColumn	DIM_DATE.Quarter (WChar)		<table border="1"> <thead> <tr> <th colspan="2">KeyColumns</th><th>(Collection)</th></tr> </thead> <tbody> <tr> <td>▪ DIM_DATE.Month (Integer)</td><td>DIM_DATE.Month (Integer)</td><td>...</td></tr> <tr> <td>▪ DIM_DATE.Year (Integer)</td><td>DIM_DATE.Year (Integer)</td><td></td></tr> <tr> <td>MemberNamesUnique</td><td>False</td><td></td></tr> <tr> <td>MembersWithData</td><td>NonLeafDataVisible</td><td></td></tr> <tr> <td>MembersWithDataCaption</td><td>MembersWithDataCaption</td><td></td></tr> <tr> <td>Name</td><td>Month</td><td></td></tr> <tr> <td>▪ NameColumn</td><td>DIM_DATE.Month_Name (WChar)</td><td></td></tr> </tbody> </table>	KeyColumns		(Collection)	▪ DIM_DATE.Month (Integer)	DIM_DATE.Month (Integer)	...	▪ DIM_DATE.Year (Integer)	DIM_DATE.Year (Integer)		MemberNamesUnique	False		MembersWithData	NonLeafDataVisible		MembersWithDataCaption	MembersWithDataCaption		Name	Month		▪ NameColumn	DIM_DATE.Month_Name (WChar)	
KeyColumns		(Collection)																																																
▪ DIM_DATE.Quarter (Integer)	DIM_DATE.Quarter (Integer)	...																																																
▪ DIM_DATE.Year (Integer)	DIM_DATE.Year (Integer)																																																	
MemberNamesUnique	False																																																	
MembersWithData	NonLeafDataVisible																																																	
MembersWithDataCaption	MembersWithDataCaption																																																	
Name	Quarter																																																	
▪ NameColumn	DIM_DATE.Quarter (WChar)																																																	
KeyColumns		(Collection)																																																
▪ DIM_DATE.Month (Integer)	DIM_DATE.Month (Integer)	...																																																
▪ DIM_DATE.Year (Integer)	DIM_DATE.Year (Integer)																																																	
MemberNamesUnique	False																																																	
MembersWithData	NonLeafDataVisible																																																	
MembersWithDataCaption	MembersWithDataCaption																																																	
Name	Month																																																	
▪ NameColumn	DIM_DATE.Month_Name (WChar)																																																	

### c. Dim\_Airport

Dimension AIRPORT phân tích dữ liệu chuyến bay theo vị trí sân bay. Dimension phân tích và so sánh hiệu suất chuyến bay, tình trạng trễ và hủy chuyến ở nhiều cấp độ không gian khác nhau, từ tổng quan theo quốc gia đến chi tiết theo từng sân bay cụ thể.

Dimension AIRPORT được xây dựng với phân cấp chiều địa lý nhằm hỗ trợ phân tích dữ liệu theo không gian: Country → State → City → Airport



Phân cấp này cho phép:

- Phân tích dữ liệu theo quốc gia, bang và thành phố
- Drill-down xuống mức chi tiết theo từng sân bay
- So sánh hiệu suất hoạt động giữa các khu vực và các sân bay cụ thể

Row Labels		Flight Count	Is Cancelled	Is Diverted	Is OTP
▪ USA					
▪ AK					
▪ Anchorage	Ted Stevens Anchorage International Airport	287	1	0	231
▪ Barrow	Wiley Post-Will Rogers Memorial Airport	16	0	1	15
▪ Bethel		15	0	0	15
▪ Cordova		10	0	0	10
▪ Deadhorse		18	0	0	16
▪ Fairbanks		36	1	0	29
▪ Gustavus		3	0	0	3
▪ Juneau		76	0	0	65
▪ Ketchikan		34	0	0	29
▪ King Salmon		3	0	0	2
▪ Kodiak		10	0	0	8
▪ Kotzebue		14	1	1	9
▪ Nome		11	0	0	9
▪ Petersburg		13	0	0	13
▪ Sitka		16	1	0	11
▪ Wrangell		13	0	0	11
▪ Yakutat		10	0	0	8
▪ AL		477	7	3	385
▪ AR		429	17	0	316
▪ AS		3	0	0	2
▪ AZ		3108	26	4	2584

Trong Dimension AIRPORT, một số thuộc tính địa lý có thể trùng tên nhưng thuộc các phạm vi khác nhau (ví dụ: các thành phố cùng tên nhưng thuộc các bang khác nhau). Để giải quyết vấn đề này, Dimension AIRPORT được cấu hình theo nguyên tắc Composite Key và Attribute Relationships, cụ thể:

- Airport được định danh bằng khóa kỹ thuật Airport Key, đảm bảo mỗi sân bay là duy nhất
- Thuộc tính City được xác định duy nhất trong phạm vi State

#### d. Dim\_Airline

Dimension AIRLINE được sử dụng nhằm phục vụ phân tích dữ liệu chuyến bay theo hãng hàng không. Dimension cho phép đánh giá và so sánh hiệu suất hoạt động của các hãng dựa trên các chỉ số như số lượng chuyến bay, tỷ lệ đúng giờ (OTP), tỷ lệ trễ và tỷ lệ hủy chuyến.

Dimension này được sử dụng ở một cấp duy nhất: Airline

Trong Dimension AIRLINE, mỗi hãng hàng không được định danh duy nhất bằng Airline Key.

Để đảm bảo hiển thị thân thiện cho người dùng trong quá trình phân tích:

- Airline Key được sử dụng làm khóa định danh kỹ thuật
- Airline Name được sử dụng làm tên hiển thị trong OLAP Cube

Row Labels	Flight Count	Is Cancelled	Is Diverted	Is OTP
American Airlines Inc.	12423	191	38	10086
Alaska Airlines Inc.	3033	9	7	2660
JetBlue Airways	4618	75	13	3576
Delta Air Lines Inc.	15051	75	32	13050
Atlantic Southeast Airlines	9843	268	27	7679
Frontier Airlines Inc.	1528	16	4	1135
Hawaiian Airlines Inc.	1303	1	2	1148
American Eagle Airlines Inc.	5058	244	12	3809
Spirit Air Lines	2013	40	1	1406
Skywest Airlines Inc.	10155	169	30	8215
United Air Lines Inc.	8918	103	24	7034
US Airways Inc.	3397	87	10	2728
Virgin America	1058	7	1	859
Southwest Airlines Co.	21602	273	61	17549
<b>Grand Total</b>	<b>100000</b>	<b>1558</b>	<b>262</b>	<b>80934</b>

### e. Dim\_Reason

Dimension REASON nhằm phục vụ phân tích dữ liệu chuyến bay theo nguyên nhân trễ, hủy hoặc chuyển hướng chuyến bay. Dimension này giúp người dùng hiểu rõ các yếu tố tác động đến hiệu suất chuyến bay, từ đó hỗ trợ việc đánh giá nguyên nhân và đề xuất các biện pháp cải thiện hoạt động vận hành.

Dimension này được sử dụng ở một cấp duy nhất: Reason

Trong Dimension REASON:

- Reason Key được sử dụng làm khóa định danh kỹ thuật, đảm bảo mỗi nguyên nhân là duy nhất
- Reason Description được sử dụng làm tên hiển thị trong OLAP Cube

## IX. MDX

### 1. Truy vấn tổng số chuyến bay theo tháng, quý, năm

```
-- Tổng số chuyến bay theo tháng
SELECT
    {[Measures].[Flight Count]} ON COLUMNS,
    NON EMPTY
    [DIM DATE].[Hierarchy].[Month].Members ON ROWS
FROM [DDS DATH]
```

Hình 1: Câu lệnh truy vấn tổng số chuyến bay theo tháng

The screenshot shows a Microsoft Analysis Services query editor window. At the top, there is a toolbar with a magnifying glass icon and a dropdown menu set to '130 %'. Below the toolbar are two tabs: 'Messages' and 'Results'. The 'Messages' tab is selected, displaying the MDX query:

```
-- Tổng số chuyến bay theo tháng
SELECT
    {[Measures].[Flight Count]} ON COLUMNS,
    NON EMPTY
    [DIM DATE].[Hierarchy].[Month].Members ON ROWS
FROM [DDS DATH]
```

The 'Results' tab is also visible. Below the tabs, there is a table with the following data:

	Flight Count
January	8103
February	7400
March	8695
April	8365
May	8569
June	8688
July	8978
August	8802
September	8016
October	8382
November	8069
December	7933

Hình 2: Kết quả truy vấn tổng số chuyến bay theo tháng

```
-- Tổng số chuyến bay theo quý
SELECT
    {[Measures].[Flight Count]} ON COLUMNS,
    NON EMPTY
        [DIM DATE].[Hierarchy].[Quarter].Members ON ROWS
FROM [DDS DATH]
```

Hình 3: Câu lệnh truy vấn tổng số chuyến bay theo quý

```
-- Tổng số chuyến bay theo quý
SELECT
    {[Measures].[Flight Count]} ON COLUMNS,
    NON EMPTY
        [DIM DATE].[Hierarchy].[Quarter].Members ON ROWS
FROM [DDS DATH]
```

	Flight Count
1	24198
2	25622
3	25796
4	24384

Hình 4: Kết quả truy vấn tổng số chuyến bay theo quý

```
-- Tổng số chuyến bay theo năm  
SELECT  
    {[Measures].[Flight Count]} ON COLUMNS,  
    NON EMPTY  
    [DIM DATE].[Hierarchy].[Year].Members ON ROWS  
FROM [DDS DATH]
```

Hình 5: Câu lệnh truy vấn tổng số chuyến bay theo năm

The screenshot shows a Microsoft SQL Server Management Studio (SSMS) interface. A query window is open with the following T-SQL code:

```
-- Tổng số chuyến bay theo năm  
SELECT  
    {[Measures].[Flight Count]} ON COLUMNS,  
    NON EMPTY  
    [DIM DATE].[Hierarchy].[Year].Members ON ROWS  
FROM [DDS DATH]
```

The code is highlighted in blue, indicating it is syntax-highlighted. Below the code, the results pane shows a single row of data:

Flight Count
2015 100000

Hình 6: Kết quả truy vấn tổng số chuyến bay theo năm

## 2. Top 5 hãng hàng không có nhiều chuyến bay nhất

```
SELECT
    { [Measures].[Flight Count] } ON COLUMNS,
    {
        TOPCOUNT(
            [DIM AIRLINE].[Airline Name].[Airline Name].MEMBERS,
            5,
            [Measures].[Flight Count]
        )
    } ON ROWS
FROM [DDS DATH]
```

117 %

	Flight Count
Southwest Airlines Co.	21602
Delta Air Lines Inc.	15051
American Airlines Inc.	12423
Skywest Airlines Inc.	10155
Atlantic Southeast Airlines	9843

Hình 7: Kết quả truy vấn top 5 hãng hàng không có nhiều chuyến bay nhất

### 3. Tỉ lệ chuyến bay đúng giờ (OTP) theo hãng hàng không

```

22  -- Tỉ lệ chuyến bay đúng giờ (OTP) theo hãng hàng không
23  WITH
24    MEMBER [Measures].[OTP Rate] AS
25      IIF(
26          [Measures].[Flight Count] = 0,
27          NULL,
28          [Measures].[Is OTP] / [Measures].[Flight Count]
29      ),
30      FORMAT_STRING = "Percent"
31
32  SELECT
33  {
34      [Measures].[OTP Rate]
35  } ON COLUMNS,
36

```

122 % ▶

	OTP Rate
All	80.21%
Alaska Airlines Inc.	86.91%
American Airlines Inc.	80.53%
American Eagle Airlines Inc.	74.67%
Atlantic Southeast Airlines	77.31%
Delta Air Lines Inc.	86.15%
Frontier Airlines Inc.	73.49%
Hawaiian Airlines Inc.	87.26%
JetBlue Airways	76.59%
Skywest Airlines Inc.	80.14%
Southwest Airlines Co.	80.34%
Spirit Air Lines	69.05%
United Air Lines Inc.	78.27%
US Airways Inc.	79.60%
Virgin America	80.25%

Hình 8: Kết quả truy vấn tỉ lệ chuyến bay đúng giờ theo hãng hàng không

#### 4. Tỉ lệ hủy chuyến theo nguyên nhân

```

WITH
--1. Tổng số chuyến bị hủy của toàn bộ nguyên nhân
MEMBER [Measures].[Total Cancellations All Reasons] AS
(
    [Measures].[Is Cancelled],
    [DIM REASON].[Reason Description].[All]
)
--2. Tỉ lệ đóng góp của từng nguyên nhân
MEMBER [Measures].[Reason Contribution Rate] AS
IIF(
    [Measures].[Total Cancellations All Reasons] = 0,

```

```

        NULL,
        [Measures].[Is Cancelled]
    /
    [Measures].[Total Cancellations All Reasons]
),
FORMAT_STRING = "0.00%"
SELECT
{
    [Measures].[Is Cancelled],
    [Measures].[Total Cancellations All Reasons],
    [Measures].[Reason Contribution Rate]
} ON COLUMNS,
NON EMPTY
FILTER(
    [DIM REASON].[Reason Description].[Reason Description].MEMBERS,
    [Measures].[Is Cancelled] > 0
    AND NOT ISEMPTY([Measures].[Is Cancelled])
    AND [DIM REASON].[Reason Description].CURRENTMEMBER.NAME <> "Not
Cancelled"
) ON ROWS
FROM [DDS DATH]
68 WITH
69 -- Tổng số chuyến BỊ HỦY của TOÀN BỘ nguyên nhân
70 MEMBER [Measures].[Total Cancellations All Reasons] AS
71 (
72     [Measures].[Is Cancelled],
73     [DIM REASON].[Reason Description].[All]
74 )
75
76
77 -- Ti lệ đóng góp của từng nguyên nhân
78
79 MEMBER [Measures].[Reason Contribution Rate] AS
80     IIF(
81         [Measures].[Total Cancellations All Reasons] = 0,

```

122 %

	Is Cancelled	Total Cancellations All Reasons	Reason Contribution Rate
Airline/Carrier	422	1558	27.09%
National Air System	259	1558	16.62%
Weather	877	1558	56.29%

Hình 9: Kết quả truy vấn tỉ lệ hủy chuyến theo nguyên nhân

## 5. Trung bình thời gian delay theo sân bay đi/đến

```
-- Trung bình delay theo sân bay đến
SELECT
    {[Measures].[Avg Arrival Delay]} ON COLUMNS,
    [DEST Airport].[Airport Name].Members ON ROWS
FROM [DDS DATH];
```

Hình 10: Câu lệnh truy vấn trung bình delay theo sân bay đến

The screenshot shows a query window with the following details:

- Code Area:**

```
43 -- Trung bình delay theo sân bay đến
44 SELECT
45     {[Measures].[Avg Arrival Delay]} ON COLUMNS,
46     [DEST Airport].[Airport Name].Members ON ROWS
47 FROM [DDS DATH];
```
- Results Area:**

	Avg Arrival Delay
All	58.84
Aberdeen Regional Airport	57.00
Abilene Regional Airport	60.00
Abraham Lincoln Capital Airport	105.29

Hình 10: Kết quả truy vấn trung bình delay theo sân bay đến

```
-- Trung bình delay theo sân bay đi
SELECT
    {[Measures].[Avg Departure Delay]} ON COLUMNS,
    [Origin Airport].[Airport Name].Members ON ROWS
FROM [DDS DATH];
```

Hình 11: Câu lệnh truy vấn trung bình delay theo sân bay đi

```

49 -- Trung bình delay theo sân bay đi
50 SELECT {[Measures].[Avg Departure Delay]} ON COLUMNS,
51 [Origin Airport].[Airport Name].Members ON ROWS
52 FROM [DDS DATH];

```

122 % ▾

Messages	Results
All	Avg Departure Delay 59.92
Aberdeen Regional Airport	416.33
Abilene Regional Airport	60.40
Abraham Lincoln Capital Airport	50.50

Hình 12: Kết quả truy vấn trung bình delay theo sân bay đi

## X. Excel report

### 1. Flight summary

SEASONAL FLIGHT PERFORMANCE REPORT 2015								
Season	Month	Total Flights	On-Time Flights	Diverted Flights	Delayed Flights	Cancelled Flights	OTP Rate	Cancellation Rate
Spring	March	8,695	6,834	24	1,861	188	78.60%	2.16%
	April	8,365	6,911	25	1,454	79	82.62%	0.94%
	May	8,569	6,939	22	1,630	91	80.98%	1.06%
Summer	June	8,688	6,550	34	2,138	158	75.39%	1.82%
	July	8,978	7,021	28	1,957	80	78.20%	0.89%
	August	8,802	7,052	27	1,750	99	80.12%	1.12%
Autumn	September	8,016	6,929	11	1,087	26	86.44%	0.32%
	October	8,382	7,344	14	1,038	41	87.62%	0.49%
	November	8,069	6,764	23	1,305	70	83.83%	0.87%
Winter	January	8,103	6,218	18	1,885	213	76.74%	2.63%
	February	7,400	5,405	16	1,995	369	73.04%	4.99%
	December	7,933	6,240	20	1,693	144	78.66%	1.82%
<b>Grand Total</b>		<b>100,000</b>	<b>80,207</b>	<b>262</b>	<b>19,793</b>	<b>1,558</b>	<b>80.21%</b>	<b>1.56%</b>

#### a. Tổng quan theo mùa

Trong năm 2015, hệ thống ghi nhận 100,000 chuyến bay, trong đó:

- 80.21% chuyến bay đúng giờ (On-Time Performance – OTP)
- 19.79% chuyến bay bị trễ
- 1.56% chuyến bay bị hủy
- 262 chuyến bị chuyển hướng (Diverted)

Hiệu suất bay có sự khác biệt rõ rệt giữa các mùa, phản ánh tác động của thời tiết, nhu cầu hành khách và áp lực khai thác.

### b. Phân tích chi tiết theo mùa

Mùa	Phân tích
Spring (March – May)	<ul style="list-style-type: none"> <li>- OTP dao động từ 78.6% – 82.62%</li> <li>- April có hiệu suất tốt nhất trong mùa xuân (82.62% OTP, cancellation chỉ 0.94%)</li> <li>- March ghi nhận tỷ lệ hủy cao hơn (2.16%), cho thấy ảnh hưởng của thời tiết chuyển mùa</li> </ul>
Summer (June – August)	<ul style="list-style-type: none"> <li>- Là mùa có lưu lượng chuyến bay cao nhất</li> <li>- OTP trung bình thấp hơn Spring và Autumn</li> <li>- June có OTP thấp nhất mùa hè (75.39%) và số delayed flights cao nhất (2,138 chuyến)</li> </ul>
Autumn (September – November)	<ul style="list-style-type: none"> <li>- Hiệu suất tốt nhất trong năm</li> <li>- OTP đạt đỉnh vào October (87.62%)</li> <li>- Tỷ lệ hủy thấp, chỉ 0.32% – 0.87%</li> </ul>
Winter (December – February)	<ul style="list-style-type: none"> <li>- Là mùa có hiệu suất thấp nhất</li> <li>- February: <ul style="list-style-type: none"> <li>• OTP chỉ 73.04%</li> <li>• Tỷ lệ hủy cao nhất năm (4.99% – 369 chuyến)</li> </ul> </li> <li>- January và December cũng ghi nhận số chuyến hủy cao</li> </ul>

### c. So sánh các chỉ số

Chỉ số	Mùa tốt nhất	Mùa kém nhất
--------	--------------	--------------

OTP Rate	Autumn	Winter
Cancellation Rate	Autumn	Winter
Delayed Flights	Autumn	Summer
Operational Stability	Autumn	Winter

#### d. Insight

- Autumn là benchmark mùa vận hành tối ưu
- Winter là điểm nghẽn lớn nhất, cần:
  - o Kế hoạch dự phòng thời tiết
  - o Buffer lịch bay
  - o Tăng cường xử lý mặt đất
- Summer cần tối ưu năng lực khai thác để giảm delay do quá tải

## 2. Airline report

TOP 5 AIRLINES PERFORMANCE & OTP REPORT

Season		Airline Name	Total Flights	OTP Rate	Avg Delay (Mins)
Spring	Summer	Southwest Airlines Co.	5,527	80.51%	25.59
		Delta Air Lines Inc.	3,833	86.93%	48.95
Month		Atlantic Southeast Airlines	2,705	76.75%	51.35
January	February	Skywest Airlines Inc.	2,601	83.51%	56.88
	Dece...	American Airlines Inc.	2,269	79.37%	46.36
<b>Grand Total</b>			<b>16,935</b>	<b>81.67%</b>	<b>40.38</b>

#### a. Tổng quan hiệu suất theo hãng

- Trong mùa Spring, hệ thống ghi nhận 16,935 chuyến bay thuộc Top 5 hãng hàng không lớn, với:
  - OTP trung bình: 81.67%
  - Thời gian trễ trung bình: 40.38 phút
- Hiệu suất giữa các hãng có sự chênh lệch rõ rệt, phản ánh khác biệt về chiến lược khai thác, quản lý đội bay và khả năng kiểm soát delay.

#### b. Phân tích chi tiết theo từng hãng

Hãng hàng không	Nhận xét
Delta Air Lines Inc.	OTP cao nhất, ưu tiên đúng giờ
Skywest Airlines Inc.	OTP tốt nhưng delay khi xảy ra thường kéo dài
Southwest Airlines Co.	Khi có delay thì thời gian delay ngắn nhất, khai thác ổn định

American Airlines Inc.	Hiệu suất trung bình, cần cải thiện OTP
Atlantic Southeast Airlines	OTP thấp nhất trong Top 5

- Nhận xét tổng quát:

- Delta dẫn đầu về OTP, cho thấy hãng hàng không kiểm soát vận hành tốt.
- Southwest nổi bật ở Avg Delay thấp, dù OTP chỉ ở mức trung bình.
- Atlantic Southeast là hãng có hiệu suất thấp nhất trong nhóm, cần tối ưu vận hành.

#### c. So sánh các chỉ số chính

Chỉ số	Hãng tốt nhất	Hãng kém nhất
OTP Rate	Delta Air Lines	Atlantic Southeast
Avg Delay	Southwest Airlines	Skywest Airlines
Operational Consistency	Southwest / Delta	Atlantic Southeast
Khả năng kiểm soát delay	Southwest	Skywest

#### d. Insight chính

- OTP cao không đồng nghĩa với delay ngắn  
→ Delta có OTP cao nhất nhưng Avg Delay vẫn khá lớn.
- Southwest theo chiến lược “delay ngắn – quay đầu nhanh”, phù hợp mô hình khai thác tần suất cao.
- Skywest & Atlantic Southeast cho thấy dấu hiệu delay kéo dài, tiềm ẩn rủi ro lan truyền chuyến bay.
- Sự khác biệt giữa các hãng cho thấy yếu tố airline-specific rất quan trọng trong bài toán dự đoán delay.

#### e. Liên hệ với mô hình dự đoán (Is\_Delayed)

- Airline Name là biến giải thích mạnh: ảnh hưởng cả xác suất delay và mức độ delay
- Có thể:
  - Encode theo historical OTP
  - Kết hợp với Season để bắt được hiệu ứng mùa – hãng
  - Các hãng có Avg Delay cao nên được xem là nhóm rủi ro trong dự báo vận hành.

### 3. Airport report

State			Airport Name	Total Flights	Avg Dep Departure Delay (Mins)
AK	AL	AR	Aberdeen Regional Airport	17	416.33
AS	AZ	CA	Abilene Regional Airport	43	60.40
CO	CT	DE	Abraham Lincoln Capital Airport	32	50.50
FL	GA	GU	Adak Airport	1	
HI	IA	ID	Akron-Canton Regional Airport	103	43.08
City			Albany International Airport	145	118.33
Green Bay			Albert J. Ellis Airport	27	178.33
Madison			Albuquerque International Sunport	343	47.53
Milwaukee			Alexandria International Airport	68	84.38
Mosinee			Alpena County Regional Airport	11	42.00
Rhinelanders			Appleton International Airport	69	83.45
Charleston			Arcata Airport	27	86.40
Casper			Arnold Palmer Regional Airport	19	60.60
Cody			Asheville Regional Airport	56	177.38
Gillette			Aspen-Pitkin County Airport	63	78.57
Jackson			Atlantic City International Airport	62	62.67
Laramie			Augusta Regional Airport (Bush Field)	51	54.17
Rock Springs			Austin-Bergstrom International Airport	755	63.09
Ithaca			Baltimore-Washington International Airport	1,606	57.67
			Bangor International Airport	9	21.00

#### a. Phân tích hiệu suất theo quy mô sân bay

Ví dụ:

- Austin-Bergstrom International Airport: 755 chuyến, Avg Departure Delay: 63.09 phút
- Baltimore–Washington International Airport: 1,606 chuyến, Avg Departure Delay: 57.67 phút

Nhận định:

- Sân bay lớn có lưu lượng cao nhưng độ trễ trung bình tương đối ổn định
- Hiệu quả vận hành tốt nhờ: Hạ tầng mạnh, Quy trình tối ưu, Nguồn lực đầy đủ

#### b. Phân tích theo sân bay nhỏ / khu vực

Ví dụ:

- Aberdeen Regional Airport: 17 chuyến, Avg Delay: 416.33 phút
- Albert J. Ellis Airport: 27 chuyến, Avg Delay: 178.33 phút

Nhận định:

- Sân bay nhỏ có biến động delay rất lớn
- Một vài chuyến trễ kéo dài có thể làm sai lệch trung bình
- Nhạy cảm với: Thời tiết, Thiếu chuyến thay thế, Hạn chế nhân sự & thiết bị

#### c. Mối quan hệ giữa lưu lượng và độ trễ

- Không có mối quan hệ tuyến tính trực tiếp giữa số chuyến và độ trễ trung bình
- Sân bay:
  - o Nhiều chuyến → Delay ổn định
  - o Ít chuyến → Delay biến động mạnh

#### 4. Root cause analysis

CANCELLATION REASONS		DELAY BREAKDOWN	
Cancel Reason	Cancelled Flights	Delay Reason	
Airline/Carrier	422	Air System Delay	9,692
National Air System	259	Airline Delay	9,725
Weather	877	Late Aircraft Delay	9,537
<b>Grand Total</b>	<b>1,558</b>	Security Delay	57
		Weather Delay	1,047

Báo cáo phân tích 1,558 chuyến bay bị hủy và gần 30,000 lượt trễ chuyến trong năm 2015, cho thấy các nguyên nhân đến từ hãng hàng không, hệ thống không lưu và yếu tố thời tiết.

##### a. Phân tích nguyên nhân hủy chuyến

- Thời tiết là nguyên nhân chiếm tỷ trọng cao nhất, hơn một nửa số chuyến hủy
- Các nguyên nhân nội bộ hãng (Airline/Carrier) đứng thứ hai
- Hệ thống không lưu quốc gia (NAS) tuy có tác động nhưng ít hơn so với thời tiết và hãng

##### b. Phân tích nguyên nhân trễ chuyến

- Airline Delay, Air System Delay và Late Aircraft Delay có mức độ tương đương nhau
- Weather Delay chiếm tỷ lệ nhỏ hơn so với các nguyên nhân vận hành
- Security Delay không đáng kể

## 5. Time analysis

### HOURLY FLIGHT PATTERNS

Day Name	Time Of Day Name	Hour	Total Flights	Avg Dep Delay (Mins)
Sunday	Early Morning	5	2,024	102.33
Monday		6	7,008	82.55
Tuesday		7	6,767	63.18
Wednesday		8	6,557	63.22
Thursday	Morning	9	6,033	56.90
Friday		10	6,407	59.50
Saturday		11	6,154	55.21
	Afternoon	12	6,108	56.99
		13	6,258	58.70
		14	5,674	59.27
		15	6,318	55.44
		16	5,735	61.43
	Evening	17	6,700	60.62
		18	5,739	60.59
		19	5,684	61.50
		20	4,448	58.88
		21	3,236	55.39
		22	2,002	54.24
		23	758	59.18
	Night	0	259	70.63
		1	91	54.20
		2	22	65.25
		3	8	54.00
		4	10	17.50
	<b>Grand Total</b>		<b>100,000</b>	<b>59.92</b>

#### a. Tổng quan theo thời điểm trong ngày

Tổng cộng ghi nhận 100,000 chuyến bay với độ trễ trung bình 59.92 phút.

Mức delay thay đổi rõ rệt theo khung giờ trong ngày, phản ánh áp lực khai thác, khả năng quay vòng tàu bay và hiệu ứng dây chuyền vận hành.

#### b. Phân tích chi tiết theo từng khung giờ

Khung giờ	Phân tích
Early Morning (05–08h)	Lượng chuyến tăng rất nhanh từ 2,024 lên đến hơn 7,000 chuyến mỗi giờ và duy trì ở mức cao đến 8h. Đây cũng là giai đoạn có delay cao nhất, đặc biệt lúc 5h đạt 102 phút – cao nhất cả ngày. Chúng tôi backlog từ cuối ngày trước và hạn chế năng lực xử lý đầu ngày.

Morning (09–11h)	Khối lượng bay ổn định ở mức khoảng 6,000 chuyến/giờ, trong khi delay giảm rõ rệt còn 55–59 phút. Đây là thời điểm hệ thống vận hành cân bằng và hiệu quả nhất.
Afternoon (12–16h)	Tiếp tục duy trì tần suất cao tương tự buổi sáng với 6,000+ chuyến mỗi giờ. Mức delay ổn định quanh 56–59 phút, ngoại trừ 16h tăng nhẹ lên 61 phút, cho thấy bắt đầu xuất hiện tích lũy chậm chuyến.
Evening (17–22h)	Lượng chuyến vẫn cao trong khung 17–19h nhưng giảm rõ rệt sau 20h. Dù lưu lượng giảm, delay duy trì khoảng 59–61 phút, cho thấy ảnh hưởng dây chuyền từ toàn bộ hoạt động ban ngày.
Night (23–04h)	Khối lượng chuyến rất thấp. Mức delay biến động mạnh (17–70 phút) do mẫu dữ liệu nhỏ, không phản ánh xu hướng vận hành. Các chuyến chủ yếu là repositioning hoặc bay ngoài giờ cao điểm.

### c. So sánh các chỉ số theo khung giờ

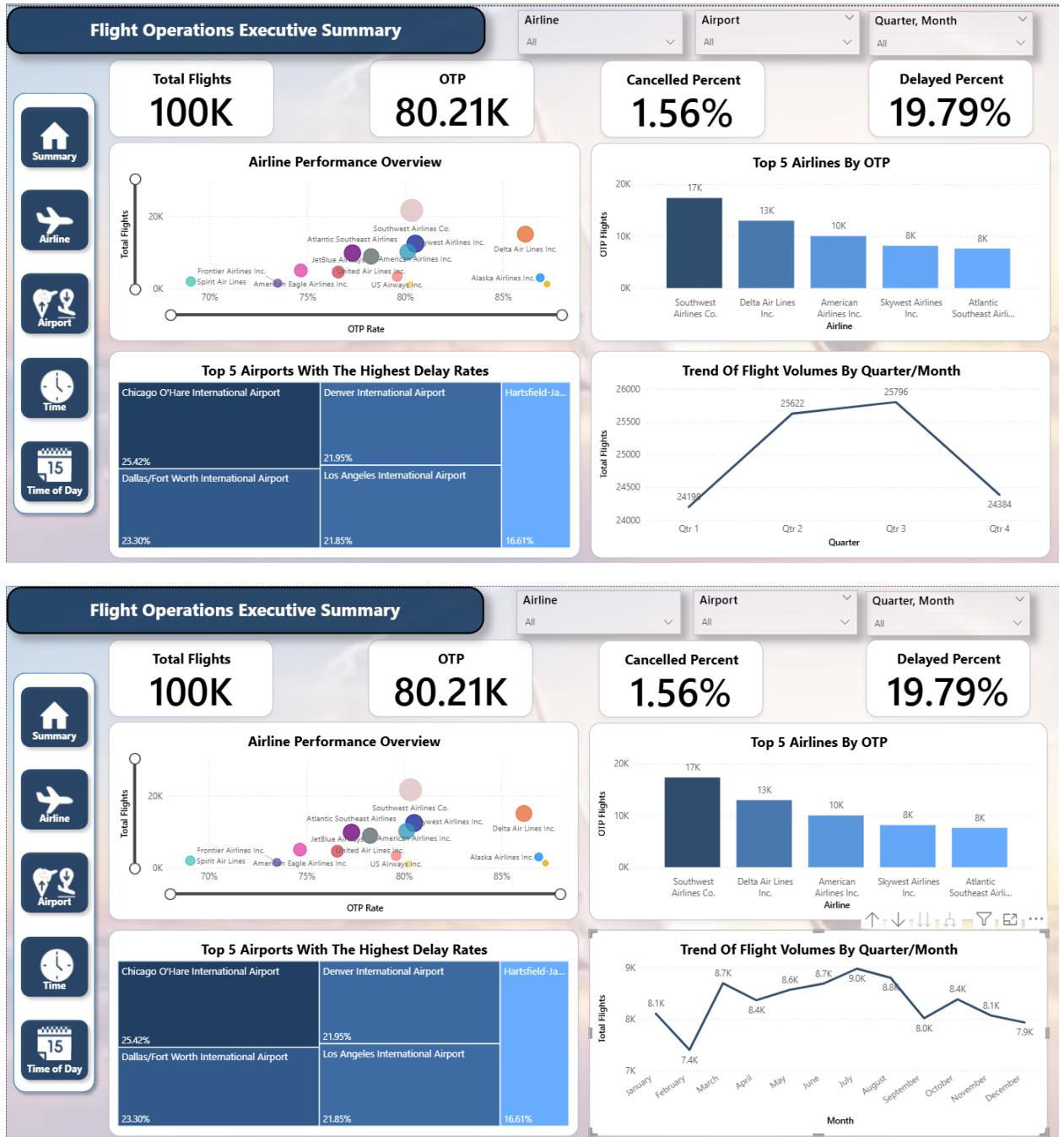
Chỉ số	Tốt nhất	Kém nhất
Delay trung bình	Morning / Afternoon	Early Morning
Ôn định vận hành	Morning	Early Morning
Khối lượng chuyến bay	Early Morning–Evening	Night
Tích lũy delay theo thời gian	Afternoon	Early Morning

### d. Insight

- Early Morning là điểm nghẽn chính: cần tăng năng lực turnaround máy bay và phân bổ slot hợp lý hơn.
- Morning và Afternoon là khung giờ vận hành tối ưu, có thể xem là benchmark năng lực khai thác.
- Evening chịu tác động dây chuyền từ các khung giờ trước; cần kiểm soát tốt kế hoạch quay đầu để tránh chậm kéo dài.
- Night không có ảnh hưởng lớn tới vận hành nhưng cần duy trì khả năng hỗ trợ xử lý backlog nếu cần.

# XI. Dashboard

## 1. Summary



Dashboard này được thiết kế nhằm phục vụ các đối tượng khác nhau, từ CEO, quản lý phòng vé, quản lý sân bay đến phòng kế hoạch chiến lược. Mỗi biểu đồ và số liệu đều có mục tiêu và insight cụ thể, giúp từng đối tượng dễ dàng phân tích và ra quyết định hiệu quả.

## **1.1. Tổng quan**

Báo cáo phân tích hiệu suất ngành hàng không năm 2015 cho thấy tổng số chuyến bay đạt **100,000** chuyến, với số lượng chuyến bay đúng giờ là **80.21K** chuyến, tỉ lệ hủy chuyến là **1.56%**, và tỉ lệ delay là **19.79%**. Những con số này phản ánh một ngành hàng không đang hoạt động hiệu quả, nhưng vẫn cần cải thiện OTP và giảm tỉ lệ delay để nâng cao chất lượng dịch vụ.

## **1.2. Phân tích chi tiết**

### **1.2.1. Hiệu suất của các hãng hàng không**

**Biểu đồ:** *Airline Performance Overview (Scatter Plot)*

**Đối tượng:** CEO hoặc quản lý cấp cao của hãng hàng không.

**Insight đạt được:**

1. Hiển thị mối quan hệ giữa số lượng chuyến bay và tỉ lệ đúng giờ (OTP):
  - Các hãng có tỉ lệ đúng giờ cao (trên 85%) thường có số lượng chuyến bay ít hơn, như Hawaiian Airlines Inc. hoặc Alaska Airlines Inc..
  - Các hãng có số lượng chuyến bay lớn (trên 15K), như Southwest Airlines Co. và Delta Air Lines Inc., thường có OTP ở mức trung bình hoặc cao (trên 80%).
2. Số lượng chuyến bay delay được sử dụng làm kích thước điểm:
  - Bubble càng lớn → số chuyến delay càng nhiều.
  - Một số hãng có bubble nổi bật càn theo dõi:
    - o Southwest Airlines Co. và American Airlines Inc.: quy mô lớn, bubble rõ rệt → áp lực vận hành dẫn tới nhiều chuyến trễ.
  - Ngược lại:
    - o Spirit Air Lines: OTP thấp (~70%) nhưng bubble nhỏ, nghĩa là ít chuyến trễ theo số lượng tuyệt đối do tổng chuyến bay thấp.
    - o Alaska Airlines Inc. và Hawaiian Airlines Inc.: bubble nhỏ, thể hiện vận hành nhát quán và ít delay.
3. Phân tích chi tiết:
  - Southwest Airlines Co. là hãng có tổng chuyến bay cao nhất (trên 21K), OTP ~80%, bubble lớn nhưng vẫn duy trì hoạt động ổn định.
  - Spirit Air Lines có OTP thấp (~70%), tuy bubble nhỏ nhưng tỷ lệ delay cao trên tổng số chuyến bay → cần chú ý cải thiện điều phổi.

- Alaska Airlines Inc. và Hawaiian Airlines Inc. có OTP cao (>85%) và bubble nhỏ, cho thấy hiệu suất tốt dù khai thác ít tuyế.

**Mục tiêu:**

1. Tăng cường dịch vụ của các hãng có OTP thấp:
  - Tập trung cải thiện cho các hãng như Spirit Air Lines và American Eagle Airlines Inc.
  - Gợi ý:
    - Tái tối ưu lịch trình
    - Cải thiện phân bổ tài nguyên tại sân bay đông đúc
    - Tăng năng lực dự báo các yếu tố gây trễ
2. Phân bổ nguồn lực phù hợp cho các hãng có số lượng chuyến bay lớn:
  - Hỗ trợ các hãng như American Airlines Inc., Southwest Airlines Co. và Delta Air Lines Inc.
  - Tập trung:
    - Giảm số chuyến bay trễ tuyệt đối
    - Tăng hiệu quả điều hành trong giờ cao điểm
    - Dùng phân tích dữ liệu để dự đoán nghẽn mạng bay

### 1.2.2. Top 5 hãng hàng không có OTP cao nhất

**Biểu đồ:** *Top 5 Airlines By OTP (Bar Chart)*

**Đối tượng:** Quản lý phòng vé hoặc các nhà phân tích dịch vụ khách hàng.

**Insight đạt được:**

- Southwest Airlines đứng đầu với OTP cao nhất (17K chuyến bay đúng giờ).
- Các hãng khác như Delta và American Airlines cũng đạt hiệu suất tốt.
- Thông tin này có thể sử dụng để tư vấn khách hàng lựa chọn hãng bay đáng tin cậy.

### 1.2.3. Tỉ lệ delay theo sân bay

**Biểu đồ:** *Top 5 Airports With The Highest Delay Rates (Treemap)*

**Đối tượng:** Quản lý sân bay hoặc các nhà điều hành vận hành.

**Insight đạt được:**

- Chicago O'Hare International Airport có tỉ lệ delay cao nhất (23.23%).
- Dallas/Fort Worth International Airport và Los Angeles International Airport cũng có tỉ lệ delay lớn.
- Điều này giúp quản lý sân bay tập trung cải thiện các yếu tố gây delay.

**Mục tiêu:** Tăng cường các biện pháp quản lý tại các sân bay có tỉ lệ delay cao.

#### 1.2.4. Xu hướng số chuyến bay theo tháng

**Biểu đồ:** *Trend of Flight Volumes By Quarter/Month (Line Chart)*

**Đối tượng:** Ban quản lý và phòng kế hoạch chiến lược.

**Insight đạt được:**

**Theo quý:**

- Số lượng chuyến bay cao nhất rơi vào quý 3 với 25.796 chuyến, trong khi các quý còn lại thấp hơn tương đối.
- Xu hướng này cho thấy quý 3 là mùa cao điểm của ngành hàng không, thường gắn liền với nhu cầu du lịch, nghỉ hè và đi lại tăng mạnh; ngược lại, các quý còn lại phản ánh giai đoạn thấp điểm hơn.
- Dữ liệu theo quý giúp doanh nghiệp hàng không lập kế hoạch phân bổ đội bay, nhân sự, lịch bay, đồng thời xây dựng chiến lược giá và marketing phù hợp cho từng giai đoạn cao điểm và thấp điểm trong năm.

**Theo tháng:**

- Số lượng chuyến bay cao nhất vào tháng 6 (8.7k) và tháng 7 (9k), thấp nhất vào tháng 2 (7.4k).
- Xu hướng này phản ánh mùa cao điểm và thấp điểm trong ngành hàng không.
- Dữ liệu này hỗ trợ lập kế hoạch nguồn lực và tiếp thị trong các giai đoạn khác nhau.

## 2. Airline:



Dashboard được thiết kế phục vụ nhiều nhóm đối tượng như ban điều hành, quản lý hàng hàng không, quản lý sân bay và phòng kinh doanh. Các biểu đồ và chỉ số cung cấp insight trọng tâm, giúp người dùng nhanh chóng nắm bắt tình hình hoạt động và hỗ trợ ra quyết định dựa trên dữ liệu.

### 2.1. Tổng quan

Báo cáo phân tích hiệu suất ngành hàng không năm 2015 cho thấy:

- Tổng số chuyến bay bị hủy: 1558 chuyến
- Tỷ lệ hủy chuyến do lỗi của hãng: 27.09%
- Thời gian delay trung bình: 35.48 phút

Các số liệu cho thấy tình trạng delay và hủy chuyến trong ngành hàng không vẫn còn ảnh hưởng đáng kể đến trải nghiệm hành khách và chi phí vận hành.

Thời gian delay trung bình trên 35 phút nhấn mạnh nhu cầu cấp thiết phải cải thiện hiệu quả vận hành và tối ưu lịch trình bay.

### 2.2. Phân tích chi tiết

#### 2.2.1. Tổng số chuyến bay bị delay

**Biểu đồ:** Flights Is Delayed (Gauge Chart)

### **Insight đạt được:**

- Tổng số chuyến bay bị delay đạt 19.79K chuyến, chiếm tỷ trọng khá đáng kể trong 100K chuyến bay.
- Việc sử dụng Gauge Chart giúp người dùng nhận biết nhanh mức độ nghiêm trọng của tình trạng delay, thay vì phải đọc số liệu chi tiết.
- Con số này cho thấy vấn đề delay không phải hiện tượng cá biệt mà là vấn đề phổ biến trong của ngành hàng không năm 2015 tại Mỹ.

### **2.2.2. Hiệu suất các hãng theo quy mô và tỉ lệ đúng giờ (OTP)**

**Biểu đồ:** Airlines By Size and OTP (Scatter Plot)

### **Insight đạt được:**

- Mối quan hệ giữa số lượng chuyến bay và OTP:
  - o Các hãng có số lượng chuyến bay thấp thường đạt OTP cao (trên 85%), cho thấy khả năng kiểm soát vận hành tốt hơn khi quy mô nhỏ.
  - o Các hãng có quy mô lớn (trên 15K chuyến) duy trì OTP ở mức trung bình - khá (~78–82%), phản ánh áp lực vận hành tăng theo quy mô.
- Kích thước điểm đại diện cho tỉ lệ trễ chuyến:
  - o Các điểm có kích thước lớn cho thấy hãng có tỉ lệ trễ cao, cần được ưu tiên theo dõi.
  - o Các điểm có kích thước lớn cho thấy hãng có tỉ lệ hủy cao, cần được ưu tiên theo dõi.
- Nhận xét:
  - o Biểu đồ cho thấy rõ sự đánh đổi giữa tăng trưởng quy mô và chất lượng dịch vụ.
  - o Đây là cơ sở quan trọng để đánh giá chiến lược mở rộng mạng bay của từng hãng.

### **2.2.3. Tổng nguyên nhân hủy chuyến theo từng hãng**

**Biểu đồ:** Total Cancellation Reason (Bar Chart)

### **Insight đạt được:**

- Late Aircraft Delay là nguyên nhân chiếm tỷ trọng cao nhất trong tổng số chuyến bị hủy ở hầu hết các hãng.
- Weather Delay đứng thứ hai, phản ánh tác động của yếu tố thời tiết lên hoạt động bay.
- Security Delay chiếm tỷ trọng nhỏ, cho thấy đây không phải nguyên nhân chính gây hủy chuyến.

### **Nhận xét:**

- Việc Late Aircraft Delay chiếm ưu thế cho thấy hiệu ứng dây chuyền trong chuỗi vận hành, khi một chuyến bay trễ kéo theo nhiều chuyến tiếp theo.
- Điều này phản ánh hạn chế trong việc tối ưu lịch trình và thời gian xoay vòng máy bay.

### **2.2.4. Số chuyến bay đúng giờ theo từng hãng**

**Biểu đồ:** On Time Flights By Airline (Line and Clustered Column Chart)

#### **Insight đạt được:**

- Các hãng có quy mô lớn (Southwest, Delta, American Airlines) chiếm số lượng chuyến đúng giờ cao hơn về mặt tuyệt đối, tuy nhiên số lượng chuyến bay delay cũng cao. Điều này cho thấy khi quy mô hoạt động tăng lên, độ phức tạp trong vận hành cũng tăng theo, làm tăng rủi ro phát sinh delay.
- Ngược lại, các hãng hàng không có quy mô nhỏ ghi nhận số lượng chuyến bay đúng giờ thấp. Tuy nhiên, số lượng chuyến bay bị delay tại các hãng này cũng ở mức rất thấp, tỉ lệ OTP cao. Điều này phản ánh lợi thế kiểm soát vận hành tốt hơn khi số lượng chuyến bay không quá lớn.
- Tuy nhiên, vẫn còn một số hãng bay có quy mô nhỏ và tỉ lệ OTP thấp. Điều này cho thấy các sân bay có quy mô nhỏ và mô hình vận hành chưa được hoàn thiện.

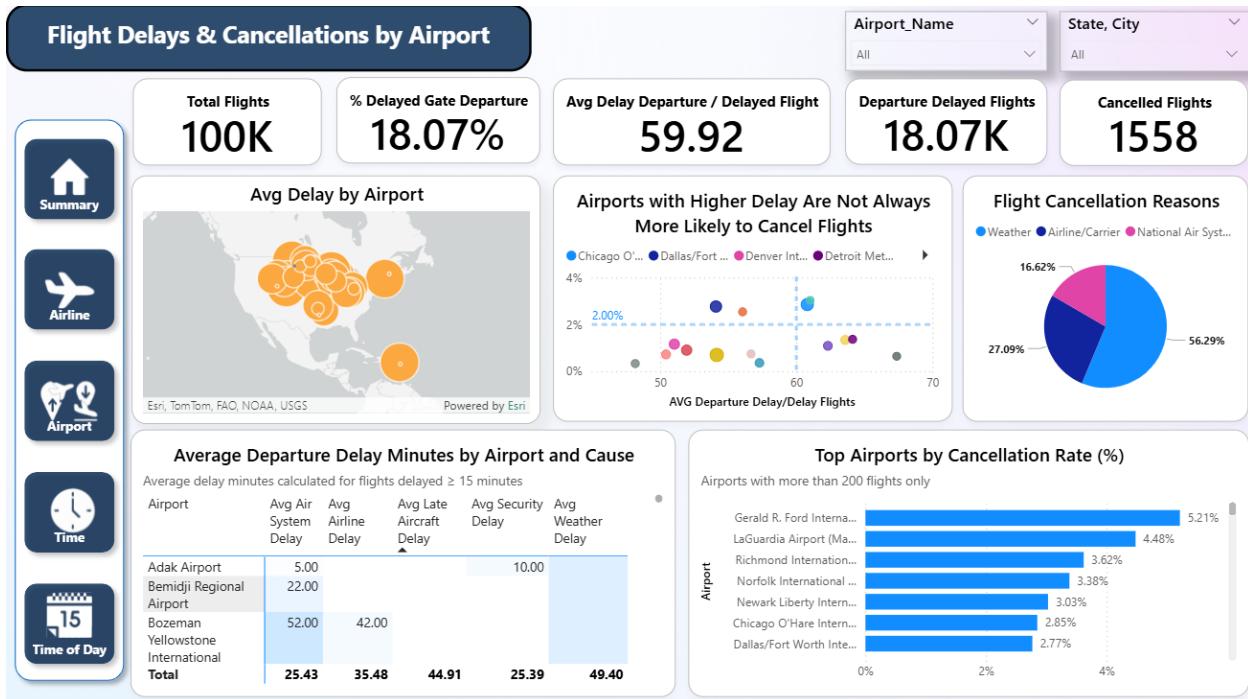
### **2.2.5. Tổng số chuyến bị hủy theo hãng**

**Biểu đồ:** Total Cancelled By Airline (Tree Map)

#### **Insight đạt được:**

- Các hãng như Southwest, Atlantic Southeast, American Eagle, American Airline,... chiếm diện tích lớn, cho thấy số lượng chuyến bay bị huỷ đáng kể.
- Các hãng có quy mô lớn thường có số chuyến bị hủy cao hơn, tuy nhiên điều này cần được đánh giá thêm theo tỉ lệ hủy, không chỉ theo số lượng tuyệt đối.

### 3. Airport:



Dashboard “Flight Delays & Cancellations by Airport” được xây dựng nhằm phân tích mức độ delay và hủy chuyến của các chuyến bay theo từng sân bay, đồng thời làm rõ các nguyên nhân chính gây ra tình trạng delay và cancellation. Thông qua việc so sánh giữa các sân bay, dashboard hỗ trợ đánh giá hiệu suất vận hành, nhận diện các điểm nóng rủi ro và cung cấp cơ sở cho các quyết định cải thiện hoạt động khai thác cũng như chiến lược kinh doanh. Dashboard phục vụ nhiều nhóm đối tượng khác nhau, bao gồm quản lý vận hành sân bay, bộ phận điều phối chuyến bay, bộ phận phòng vé và chăm sóc khách hàng, cũng như hành khách có nhu cầu lựa chọn sân bay hoặc hãng bay có mức độ rủi ro delay thấp.

#### 3.1. Tổng quan:

Báo cáo phân tích tình trạng delay và hủy chuyến theo sân bay trong năm 2015 cho thấy:

- Tổng số chuyến bay được phân tích: khoảng 100.000 chuyến
- Tỷ lệ chuyến bay khởi hành bị delay từ công từ 15 phút trở lên là khoảng 18%, phản ánh mức độ phổ biến của tình trạng trễ chuyến trong hoạt động khai thác tại các sân bay
- Thời gian delay khởi hành trung bình của các chuyến bị delay vào khoảng 60 phút, cho thấy khi delay xảy ra thì mức độ ảnh hưởng là tương đối lớn, không chỉ mang tính ngắn hạn.

- Tổng số chuyến bay bị hủy: 1.558 chuyến

Các số liệu cho thấy delay là vấn đề phổ biến và có mức độ ảnh hưởng lớn, trong khi hủy chuyến xảy ra với tần suất thấp hơn nhưng vẫn gây tác động đáng kể đến trải nghiệm hành khách và hoạt động khai thác.

Thời gian delay trung bình gần 60 phút cho thấy nhu cầu cần thiết phải cải thiện hiệu quả vận hành tại các sân bay, đặc biệt là trong công tác điều phối chuyến bay và quản lý nguyên nhân gây delay.

### **3.2. Phân tích chi tiết:**

#### **3.2.1. Tổng quan hiệu suất Delay & Cancellation theo sân bay**

**Biểu đồ:** KPI Summary Cards (Total Flights, % Delayed Gate Departures, Avg Departure Delay per Delayed Flight (min), Departure Delayed Flights, Cancelled Flights)

##### **Insight đạt được:**

- Delay là vấn đề phổ biến trong hệ thống bay năm 2015, với khoảng 18 % các chuyến bay khởi hành bị trễ từ cổng (gate) từ 15 phút trở lên, cho thấy delay không phải là các trường hợp đơn lẻ mà mang tính hệ thống trong hoạt động khai thác. Khi drill-down theo từng sân bay, tỷ lệ delay có thể cao hơn mức trung bình, phản ánh sự khác biệt về điều kiện và năng lực vận hành giữa các sân bay.
- Mức độ delay khi đã xảy ra tương đối nghiêm trọng, với thời gian delay trung bình gần 60 phút cho mỗi chuyến bị delay. Tại một số sân bay, mức delay trung bình tăng cao do chịu ảnh hưởng mạnh từ các yếu tố như thời tiết hoặc hiện tượng late aircraft.
- Tỷ lệ hủy chuyến thấp hơn đáng kể so với delay, cho thấy các hãng hàng không có xu hướng chấp nhận delay để duy trì chuyến bay. Tuy nhiên, khi phân tích theo từng sân bay, vẫn tồn tại một số điểm có tỷ lệ hủy cao hơn mặt bằng chung, phản ánh mức độ rủi ro vận hành khác nhau giữa các sân bay.

**Mục tiêu:** Cung cấp cái nhìn tổng quan về tình trạng delay và cancellation, làm cơ sở cho việc xác định mức độ ưu tiên cải thiện vận hành và lựa chọn trọng tâm giữa giảm delay hoặc giảm hủy chuyến.

#### **3.2.2. Phân bố thời gian delay trung bình theo sân bay**

**Biểu đồ:** Avg Delay by Airport (Map)

**Đối tượng:** Quản lý sân bay và bộ phận điều phối chuyến bay

##### **Insight đạt được:**

- Thời gian delay trung bình phân bố không đồng đều giữa các sân bay.
  - Mức delay trung bình khác biệt rõ rệt trên toàn nước Mỹ, với một số sân bay cao hơn đáng kể.

- Điều này cho thấy delay mang tính cục bộ, chịu ảnh hưởng lớn từ điều kiện và năng lực vận hành của từng sân bay, không chỉ do đặc thù chung của ngành.
- Sự khác biệt rõ rệt theo khu vực địa lý
  - Khu vực East Coast và Central US có mật độ delay cao hơn so với nhiều sân bay ở miền Tây.
  - Nguyên nhân có thể đến từ thời tiết phức tạp, tần suất bay cao và sự phụ thuộc giữa các sân bay trong khu vực.

➔ Delay có xu hướng mang tính “lan truyền theo vùng”, đặc biệt tại các khu vực có mật độ sân bay và chuyến bay dày đặc.

**Mục tiêu:** Xác định các điểm nóng delay theo không gian địa lý, hỗ trợ nhận diện rủi ro vận hành, ưu tiên cải thiện tại các hub lớn và cung cấp thông tin tham khảo cho phòng vé, chăm sóc khách hàng và hành khách khi lựa chọn sân bay có mức delay thấp hơn.

### 3.2.3. Phân bố thời gian delay trung bình theo sân bay

**Biểu đồ:** Airports with Higher Delay Are Not Always More Likely to Cancel Flights (Scatter Plot)

**Đối tượng:** CEO chiến lược và Business Analyst

**Insight đạt được:**

- Delay cao không đồng nghĩa với tỉ lệ hủy cao
  - Scatter plot cho thấy các điểm dữ liệu phân bố rải rác quanh hai trục tham chiếu, không tạo thành xu hướng tuyến tính rõ ràng.
  - Có nhiều sân bay nằm bên phải đường  $x = 60$  phút (delay cao) nhưng dưới đường  $y = 2\%$  (tỉ lệ hủy thấp).
- ➔ Một số sân bay có khả năng xử lý tình huống tốt, chấp nhận delay kéo dài thay vì hủy chuyến.
- Đường tham chiếu  $x = 60$  phút đóng vai trò ngưỡng nghiệp vụ để phân biệt delay chấp nhận được và delay kéo dài ảnh hưởng đến hành khách; các sân bay nằm bên phải đường này được xem là có rủi ro delay cao và cần ưu tiên cải thiện vận hành.
- ➔ Không phải tất cả sân bay có delay trên 60 phút đều có tỷ lệ hủy cao, cho thấy sự khác biệt trong chiến lược vận hành.

- Đường tham chiếu  $y = 2\%$  đóng vai trò ngưỡng cảnh báo về tỷ lệ hủy chuyến; các sân bay nằm trên mức này được xem là có tỷ lệ hủy cao hơn trung bình, gây tác động tiêu cực đến trải nghiệm hành khách và doanh thu.
  - ➔ Một số sân bay có tỷ lệ hủy cao dù mức delay trung bình không quá lớn, cho thấy cancellation không chỉ phụ thuộc vào delay.
- Phân nhóm sân bay dựa trên hai đường tham chiếu:
  - Delay thấp – Cancel thấp (Góc dưới trái): Hoạt động hiệu quả, vận hành ổn định.
  - Delay cao – Cancel thấp (Góc dưới phải): Chấp nhận delay để duy trì chuyến bay, khả năng phục hồi tốt.
  - Delay thấp – Cancel cao (Góc trên trái): Dễ hủy chuyến khi gặp sự cố, cần xem xét lại chính sách vận hành.
  - Delay cao – Cancel cao (Góc trên phải): Nhóm rủi ro cao, cần ưu tiên cải thiện cả delay và cancellation.
- ➔ Việc phân nhóm này hỗ trợ xác định ưu tiên cải thiện khác nhau cho từng nhóm sân bay.

**Mục tiêu:** Phân tích mối quan hệ giữa delay trung bình và tỷ lệ hủy chuyến theo sân bay, cho thấy delay cao không đồng nghĩa với hủy chuyến cao, qua đó hỗ trợ phân loại rủi ro vận hành và xác định ưu tiên cải thiện phù hợp.

### 3.2.4. Phân tích nguyên nhân hủy chuyến bay

**Biểu đồ:** Flight Cancellation Reasons (Pie Chart)

**Đối tượng:** CEO, Bộ phận quản lý rủi ro, Phòng kế hoạch

**Insight đạt được:**

- Thời tiết là nguyên nhân hủy chuyến lớn nhất, cho thấy cancellation chịu tác động mạnh từ yếu tố khách quan khó kiểm soát.
- Nguyên nhân từ hãng bay (Airline/Carrier) vẫn chiếm tỷ trọng đáng kể, phản ánh vai trò của lập lịch, quản lý đội bay và vận hành nội bộ.
- Nguyên nhân từ hệ thống hàng không quốc gia (National Air System) chiếm tỷ trọng thấp hơn nhưng vẫn ảnh hưởng đến hủy chuyến tại một số sân bay.
- ➔ Cơ cấu nguyên nhân hủy chuyến khác nhau giữa các sân bay; phân tích theo từng sân bay giúp nhận diện vấn đề cục bộ và đề xuất giải pháp phù hợp.

**Mục tiêu:** Phân biệt nguyên nhân khách quan và chủ quan gây hủy chuyến để xác định phạm vi kiểm soát và ưu tiên giải pháp cải thiện vận hành cho từng sân bay.

### 3.2.5. Phân tích chi tiết nguyên nhân delay theo từng sân bay

**Biểu đồ:** Average Arrival Delay Minutes by Airport and Cause (Table)

**Đối tượng:** Operation Manager, Data Analyst

### **Insight đạt được:**

- Late Aircraft là nguyên nhân gây delay trung bình cao nhất tại nhiều sân bay, cho thấy tính dây chuyền của delay giữa các chuyến bay.
  - Weather Delay duy trì ở mức cao và khá ổn định, phản ánh tác động phổ biến của thời tiết khi xảy ra delay.
  - Security Delay ít xảy ra nhưng có thể gây delay rất lớn tại một số sân bay (như George Bush Intercontinental), cho thấy mức độ ảnh hưởng nghiêm trọng khi phát sinh.
  - Airline Delay và Air System Delay có mức delay trung bình thấp hơn, thường mang tính vận hành nội bộ và gây delay ngắn hơn.
- ➔ Thời gian delay khác nhau đáng kể theo từng nguyên nhân và sân bay; phân tách theo nguyên nhân giúp xác định yếu tố tác động lớn nhất và ưu tiên giải pháp phù hợp.

**Mục tiêu:** Phân tích và so sánh thời gian delay trung bình theo từng nguyên nhân tại các sân bay để xác định nguyên nhân gây delay nghiêm trọng nhất và định hướng ưu tiên cải thiện vận hành.

### **3.2.6. Xếp hạng sân bay theo tỉ lệ hủy chuyến**

**Biểu đồ:** Top Airports by Cancellation Rate (%) (Bar Chart)

**Đối tượng:** CEO phòng vé, Bộ phận chăm sóc khách hàng,

### **Insight đạt được:**

- Tỷ lệ hủy chuyến khác biệt rõ rệt giữa các sân bay, ngay cả khi chỉ xét các sân bay có lưu lượng khai thác tương đối lớn (trên 200 chuyến bay). Sân bay có tỷ lệ hủy cao nhất đạt khoảng 5,21%, trong khi các sân bay còn lại trong nhóm dẫn đầu dao động trong khoảng 2,77%–4,48%, cho thấy mức độ ổn định vận hành không đồng đều giữa các sân bay.
- Các sân bay lớn và sân bay trung chuyển như LaGuardia, Chicago O'Hare hay Dallas/Fort Worth thường xuất hiện trong nhóm có tỷ lệ hủy cao, do chịu áp lực điều phối và lưu lượng chuyến bay lớn.
- Biểu đồ giúp nhanh chóng nhận diện các sân bay có tỷ lệ hủy chuyến cao, làm cơ sở cho các bộ phận kinh doanh và chăm sóc khách hàng chủ động tư vấn hành khách, đồng thời hỗ trợ bộ phận vận hành xác định các điểm cần ưu tiên theo dõi và cải thiện.

**Mục tiêu:** Xác định và so sánh các sân bay có tỷ lệ hủy chuyến cao nhất để hỗ trợ ra quyết định vận hành và kinh doanh, đặc biệt trong công tác tư vấn khách hàng và tối ưu chính sách bán vé.

### **3.3. Cơ chế lọc và tương tác dashboard:**

Dashboard hỗ trợ các slicer Airport và State, City, cho phép người dùng lọc theo sân bay, bang hoặc thành phố cụ thể. Khi lựa chọn thay đổi, toàn bộ KPI và biểu đồ sẽ được cập nhật động, phản ánh chính xác tình trạng delay và hủy chuyến trong

phạm vi đã chọn. Nhờ đó, người dùng có thể linh hoạt chuyển đổi giữa góc nhìn tổng thể toàn hệ thống và phân tích chi tiết theo từng sân bay hoặc khu vực địa lý.

#### 4. Time-delay:



#### Mục tiêu chính của dashboard là:

Dashboard được thiết kế nhằm cung cấp một bức tranh toàn cảnh về hiệu suất vận hành bay dưới lăng kính thời gian. Mục tiêu cốt lõi là chuyển đổi dữ liệu thành các thông tin chiến lược (insights) để giải quyết bài toán tối ưu hóa vận hành, cụ thể:

- Nhận diện tính chu kỳ (Seasonality & Trends): Làm rõ xu hướng trễ chuyến biến động như thế nào theo thời gian (mùa, tháng và các ngày trong tuần) để nâng cao khả năng dự báo.
- Tối ưu hóa nguồn lực (Resource Optimization): Xác định chính xác các “khung giờ nóng” (hotspots) về delay, từ đó hỗ trợ doanh nghiệp điều phối nhân sự (mặt đất, phi hành đoàn) và nguồn lực kỹ thuật hợp lý, tránh lãng phí hoặc quá tải.
- Nâng cao chỉ số đúng giờ (OTP Improvement): Phân tích nguyên nhân gốc rễ và đánh giá hiệu suất của từng hãng hàng không nhằm đưa ra các quyết định cải tiến quy trình hoặc điều chỉnh chiến lược hợp tác.

#### Cụ thể có thể trả lời các câu hỏi chiến lược sau:

1. Tỷ lệ delay toàn hệ thống đang ở mức báo động hay chấp nhận được?
2. Mùa, tháng và ngày nào trong tuần là “điểm nóng” của việc trễ chuyến?

3. Hãng hàng không nào đang có hiệu suất vận hành kém nhất (về cả tần suất và thời gian delay)?
4. Nguyên nhân nào đóng góp nhiều nhất vào tổng thời gian chờ đợi của khách hàng?

**Dashboard hướng đến phục vụ bộ phận quản lý và vận hành (Operations Management) của doanh nghiệp hàng không/sân bay:**

- Nhà quản lý vận hành: Cần theo dõi các điểm nóng delay để xử lý sự cố tức thời và lập kế hoạch dài hạn.
- Bộ phận kiểm soát chất lượng: Cần số liệu để đánh giá KPI của các hãng hàng không và tìm ra các quy trình gây delay.
- (*Giá trị gia tăng*) Khách hàng/Đại lý vé: Gián tiếp hưởng lợi thông qua việc tham khảo dữ liệu để lựa chọn khung giờ bay có rủi ro trễ thấp nhất.

#### **4.1. Tổng quan**

**KPI Summary Cards** (Flight Count, Delayed Flights, Percent Delay, Avg Arrival Delay).

**Mục tiêu:**

- Cung cấp bức tranh toàn cảnh về “sức khỏe” vận hành.
- Đặt ra mức tham chiếu (baseline) để so sánh khi drill-down vào từng hãng bay hoặc mùa vụ/tháng.

**Insight đạt được:**

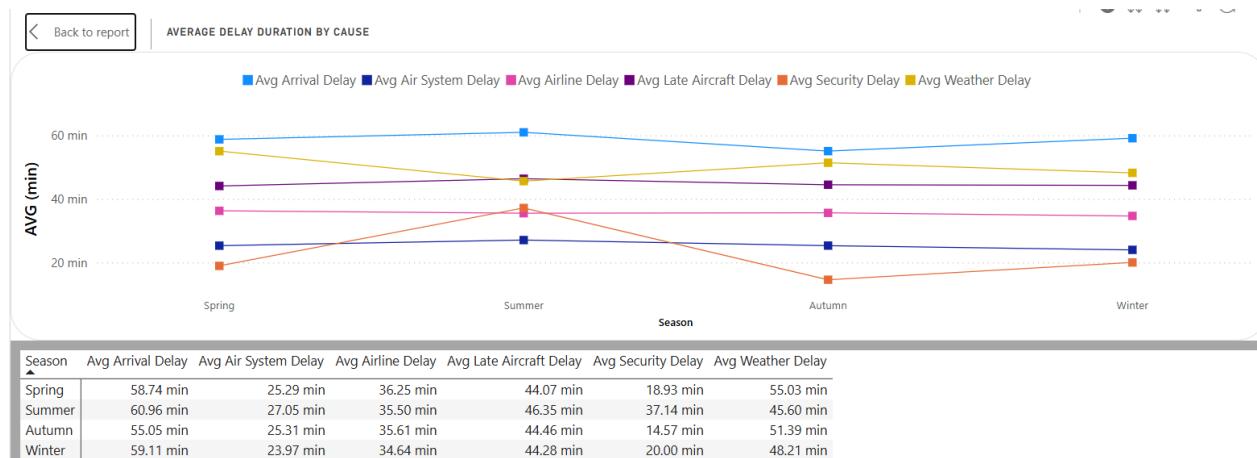
Tổng số chuyến bay được khảo sát là 100,000 chuyến, trong đó ghi nhận 19,79k chuyến bị trễ, tương ứng với tỷ lệ delay là 19,79%. Đặc biệt, thời gian trễ trung bình mỗi chuyến lên tới 58.84 phút. Những con số này phản ánh một hệ thống vận hành đang chịu áp lực lớn về mặt thời gian, với tần suất sự cố khá cao và khả năng phục hồi chậm, đòi hỏi các biện pháp can thiệp quyết liệt để giảm thiểu thời gian chờ đợi và nâng cao chất lượng dịch vụ.

## 4.2. Phân tích xu hướng delay và nguyên nhân delay theo mùa, tháng

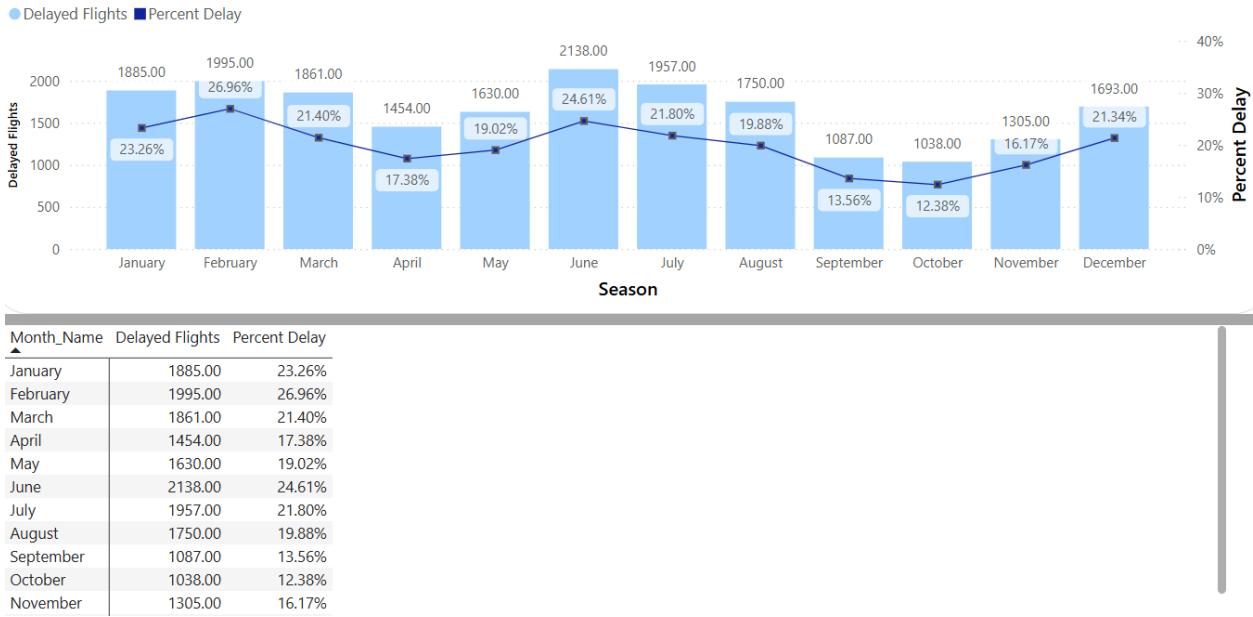
### 4.2.1. Phân tích xu hướng, tính mùa vụ và nguyên nhân



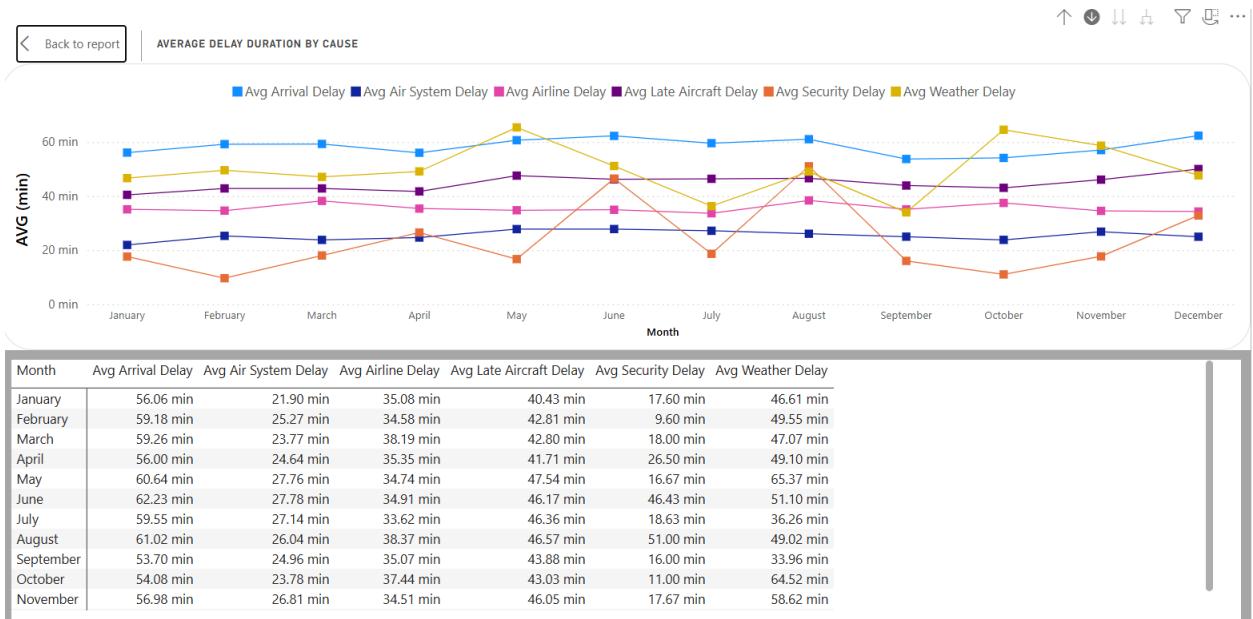
Xu hướng số chuyến bay delay và tỉ lệ delay theo mùa



Xu hướng thời gian delay trung bình theo mùa của từng nguyên nhân cụ thể



Xu hướng số chuyến bay delay và tỉ lệ delay theo tháng



Xu hướng thời gian delay trung bình theo mùa của từng nguyên nhân cụ thể

Dựa vào biểu đồ, dữ liệu cho thấy sự phân hóa rõ rệt về hiệu suất giữa các giai đoạn trong năm, định hình hai thái cực vận hành:

### Giai đoạn cao điểm (High Season Challenges):

- Mùa Hè (Summer):** Là giai đoạn vận hành kém hiệu quả nhất. Tháng 6 ghi nhận đỉnh điểm của năm với tỷ lệ trễ 23,21% và thời gian trễ trung bình 62,23

- phút. Dù tỷ lệ trễ giảm dần vào tháng 7 (21,09%) và tháng 8 (18,97%), thời gian chờ đợi vẫn duy trì ở mức cao trên 60 phút.
- Mùa Đông (Winter):** Duy trì tỷ lệ trễ cao tương đương mùa Hè (21,34%), đặc biệt tháng 2 có tỷ lệ trễ lên tới 23,13%.

### Giai đoạn Thấp điểm (Low Season Opportunities):

- Mùa Thu (Autumn):** Là giai đoạn ổn định nhất với tỷ lệ trễ trung bình chỉ 13,53%. Tháng 10 đạt hiệu suất tối ưu nhất năm (tỷ lệ trễ 11,95%). Tuy nhiên, xu hướng trễ bắt đầu tăng trở lại vào Tháng 11 (15,44%), báo hiệu sự chuyển giao sang cao điểm cuối năm.

#### 4.2.2. Phân tích nguyên nhân cụ thể

Việc đối chiếu dữ liệu tần suất và nguyên nhân trễ chuyến làm rõ các yếu tố tác động chính:

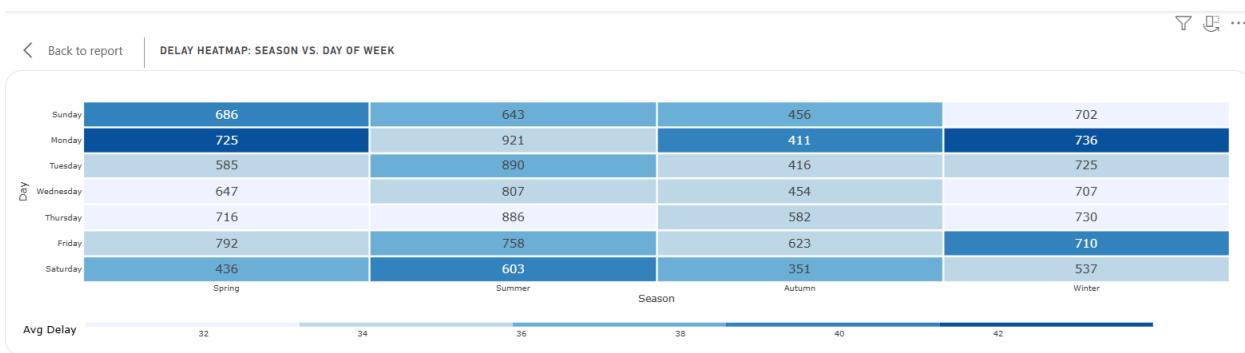
- Vấn đề về năng lực an ninh (Security Capacity):**
  - Có mối tương quan thuận giữa cao điểm du lịch và thời gian trễ do an ninh. Cụ thể, chỉ số *Security Delay* tăng vọt vào Tháng 6 (46,43 phút) và Tháng 8 (51,00 phút), trong khi giảm sâu vào Tháng 7 và các tháng thấp điểm.
  - Kết luận:* Đây là nút thắt cổ chai về nhân sự/quy trình tại các thời điểm lưu lượng khách tăng đột biến.
- Vấn đề hệ thống về lịch bay (Systemic Scheduling Issues):**
  - Chỉ số *Late Aircraft Delay* (Tàu bay về muộn) duy trì đường xu hướng đi ngang và ổn định ở mức cao (40 - 47 phút) xuyên suốt 12 tháng, không phụ thuộc vào mùa vụ.
  - Kết luận:* Đây là nguyên nhân mang tính cấu trúc, phản ánh việc sắp xếp lịch bay quá dày đặc, thiếu thời gian đệm (buffer time) để xử lý quay đầu tàu bay.
- Biến động do Thời tiết (Weather Volatility):**
  - Mặc dù tần suất biến động, mức độ nghiêm trọng (thời gian chờ) của *Weather Delay* lại đạt đỉnh vào các tháng chuyển mùa như Tháng 5 và Tháng 10 (đều trên 60 phút), dù tháng 10 có tổng tỷ lệ trễ thấp nhất.

#### 4.2.3. Đề xuất

Dựa trên các phân tích trên, đề xuất các hành động cụ thể:

- Tối ưu hóa nguồn lực an ninh: Tập trung tăng cường nhân sự và mở rộng năng lực soi chiếu vào Tháng 6 và Tháng 8 để giải quyết tình trạng tắc nghẽn cục bộ đã được nhận diện.
- Điều chỉnh cấu trúc lịch bay: Xem xét nói rộng thời gian quay đầu (turnaround time) để giảm thiểu chỉ số *Late Aircraft Delay*, vốn đang là nguyên nhân gây trễ dai dẳng nhất trong năm.
- Chiến lược mùa thấp điểm: Tận dụng khoảng thời gian Tháng 9 và Tháng 10 để thực hiện bảo trì lớn hoặc đào tạo nhân sự, nhưng cần kích hoạt kế hoạch cao điểm ngay từ giữa Tháng 11 khi xu hướng trễ bắt đầu tăng trở lại.

### 4.3. Phân tích xu hướng delay theo tuần



*Phân tích xu hướng số chuyến delay và trung bình thời gian delay theo tuần*

#### 4.3.1. Tổng quan xu hướng

Dữ liệu phân bổ trễ chuyến theo ngày trong tuần (Day of Week) chỉ ra quy luật vận hành rõ rệt: áp lực tập trung lớn nhất vào đầu tuần (Thứ Hai) và giảm sâu vào cuối tuần (Thứ Bảy). Đặc biệt, Mùa Hè không chỉ cao điểm vào đầu tuần mà duy trì áp lực cảng thẳng diện rộng gần như toàn thời gian, trong khi Mùa Thu cho thấy sự dư thừa năng lực đáng kể.

#### 4.3.2. Phân tích nhịp độ vận hành tuần

Dựa trên số lượng chuyến bay bị trễ, nhận diện 3 điểm nhấn chính trong chu kỳ tuần:

##### - Thứ Hai – Ngày cao điểm:

- o Là ngày ghi nhận số lượng chuyến trễ cao nhất hoặc nhì tại hầu hết các mùa. Điểm đỉnh là Thứ Hai của Mùa Hè với 921 chuyến trễ - con số cao nhất trong toàn bộ bảng dữ liệu.

- *Kết luận:* Đây là hệ quả của việc dồn ứ khách công tác (business travelers) và tồn đọng sự cố từ Chủ Nhật, tạo áp lực lớn nhất lên quy trình check-in và kiểm soát không lưu.

- **Thứ Bảy – Ngày thấp điểm:**

- Ngược lại, Thứ Bảy là ngày có hiệu suất tốt nhất tuần với số lượng trễ thấp kỷ lục. Cụ thể, Thứ Bảy Mùa Thu chỉ có 351 chuyến trễ, và Mùa Xuân là 436 chuyến.
- *Kết luận:* Áp lực bầu trời giảm mạnh vào ngày này, tạo cơ hội lý tưởng cho việc điều phối lại máy bay hoặc cho nhân sự nghỉ ngơi.

- **Thứ Sáu - Áp lực cuối tuần:**

- Dữ liệu ghi nhận xu hướng tăng vọt cục bộ vào Thứ Sáu tại Mùa Xuân (792 chuyến) và Mùa Đông (710 chuyến).
- *Kết luận:* Phản ánh đặc thù nhu cầu du lịch ngắn ngày hoặc về quê cuối tuần trong các mùa lễ hội/mát mẻ.

### **4.3.3. Phân tích tác động mùa vụ lên cấu trúc tuần**

Sự khác biệt về cấu trúc tuần giữa mùa hè và mùa thu là rất lớn:

- Mùa hè: Duy trì số lượng trễ rất cao ở cả các ngày giữa tuần như thứ ba (890), thứ năm (886). Hệ thống gần như không có thời gian phục hồi (recovery time) giữa các ngày.
- Mùa thu: Ngày cao điểm nhất của mùa thu (thứ hai: 411 chuyến) vẫn thấp hơn đáng kể so với ngày thấp điểm nhất của mùa hè (thứ bảy: 603 chuyến).

### **4.3.4. Kiến nghị quản trị**

Từ các phát hiện trên, đề xuất phương án tối ưu hóa:

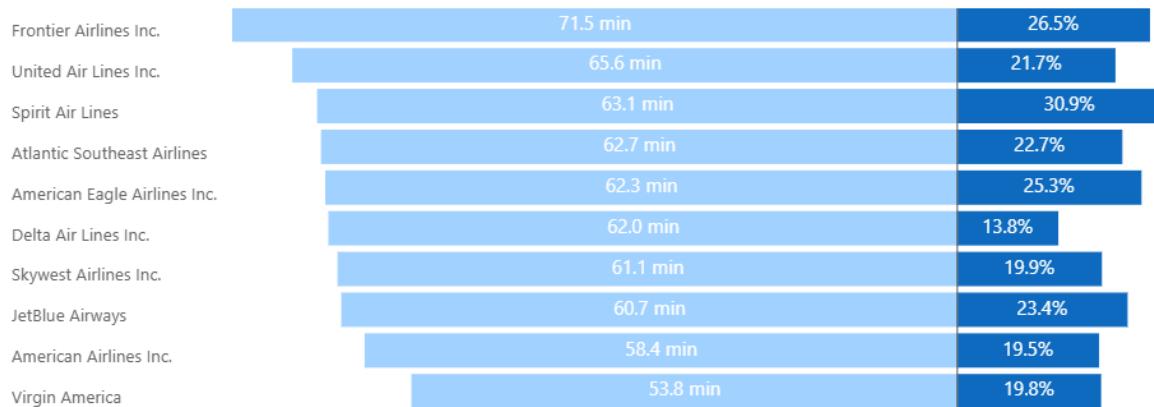
- **Tối ưu hóa lịch trực (Smart Rostering):**

- Thiết lập chế độ trực 100% quân số (không nghỉ phép) vào các ngày thứ hai (đặc biệt là mùa hè) và thứ sáu (mùa xuân/đông).
- Ngược lại, cắt giảm định biên nhân sự hoặc bố trí nghỉ bù tập trung vào Thứ Bảy để tối ưu chi phí vận hành.

- **Chiến lược thương mại (Commercial Strategy):**

- Đẩy mạnh các chương trình khuyến mãi vé rẻ cho các chuyến bay ngày thứ bảy để điều hướng khách hàng (Demand Shifting), giúp san sẻ tải trọng cho ngày Thứ Hai và Thứ Sáu đang quá tải.
- Tận dụng “vùng trũng” Mùa Thu để thực hiện các chiến dịch kích cầu doanh nghiệp hoặc du lịch MICE vào giữa tuần.

#### 4.4. Phân tích hiệu suất vận hành theo hãng hàng không



*Biểu đồ xếp hạng hàng hàng không theo tỉ lệ delay và trung bình công delay.*

Dữ liệu xếp hạng các hãng hàng không cho thấy sự phân hóa rõ rệt về năng lực kiểm soát trễ chuyến. Mặc dù thời gian trễ trung bình của toàn ngành dao động ở mức cao (từ 33.06 phút đến 72 phút), tỷ lệ trễ chuyến (tần suất) lại có biên độ dao động rất lớn, từ mức thấp 12.67% đến mức cao 29.55%. Điều này phản ánh sự chênh lệch về quy trình quản lý vận hành và khả năng tuân thủ lịch bay giữa các hãng.

Dựa trên hai chỉ số *thời gian trễ trung bình* (Duration) và *tỷ lệ trễ* (Frequency), có thể thấy:

- Frontier Airlines Inc.: Là đơn vị có chỉ số rủi ro cao nhất về mặt thời gian, với mức trễ trung bình kỷ lục 72 phút/chuyến. Đồng thời, tỷ lệ trễ cũng ở mức rất cao (25.73%).
- Spirit Air Lines: Là đơn vị có tần suất trễ cao nhất trong danh sách, lên tới 29.55%. Mặc dù thời gian trễ thấp hơn Frontier (63 phút), nhưng với việc gần 1/3 số chuyến bay gặp sự cố, đây là hãng có độ ổn định thấp nhất.
- Hawaiian Airline Inc: Ghi nhận hiệu suất vận hành tốt nhất về mặt tần suất với tỷ lệ trễ chỉ 12.67%, thấp hơn đáng kể so với mức trung bình ngành. Thời gian trung bình delay là 33.06 cũng thấp nhất toàn ngành.

## 5. Time of Day – Delay

Dashboard “Delay Patterns By Time Of Date” được xây dựng nhằm phân tích chuyên sâu các mô hình và xu hướng trễ chuyến (delay) dựa trên yếu tố thời gian (khung giờ trong ngày và các thứ trong tuần), từ đó làm rõ sự biến động của hiệu suất bay theo chu kỳ hoạt động.

Qua việc trực quan hóa mối tương quan giữa tỷ lệ phần trăm delay, thời gian trễ trung bình và lưu lượng chuyến bay bị hoãn theo từng thời điểm (Sáng sớm, Sáng, Chiều, Tối, Đêm), dashboard hỗ trợ đánh giá tác động của yếu tố thời gian lên vận hành, nhận diện các khung giờ "cao điểm" có rủi ro cao và cung cấp cơ sở dữ liệu quan trọng để điều chỉnh lịch bay hoặc phân bổ nguồn lực mặt đất hợp lý.

Dashboard phục vụ nhiều nhóm đối tượng khác nhau, bao gồm bộ phận lập kế hoạch bay, quản lý vận hành sân bay, đội ngũ kiểm soát rủi ro, cũng như hành khách có nhu cầu tối ưu hóa lịch trình di chuyển để tránh các khung giờ thường xuyên xảy ra sự cố.

### 5.1. Tổng quan

Báo cáo phân tích delay theo thời điểm trong ngày cho thấy trễ chuyến không phân bố đồng đều mà chịu ảnh hưởng mạnh bởi nhịp độ khai thác và chu kỳ vận hành trong ngày. Dữ liệu phản ánh sự khác biệt rõ rệt giữa các khung giờ về cả tỷ lệ trễ chuyến (Percent Delay) và thời gian chờ trung bình (Avg Airline Delay), cho thấy delay là hiện tượng mang tính hệ thống theo thời gian, không mang tính ngẫu nhiên.

Đáng chú ý, mỗi quan hệ giữa tần suất trễ và mức độ trễ không tỷ lệ thuận, hàm ý rằng các chiến lược vận hành cần được thiết kế khác nhau cho từng khung giờ trong ngày.

### 5.2. Tổng quan hiệu suất Delay theo khung giờ trong ngày

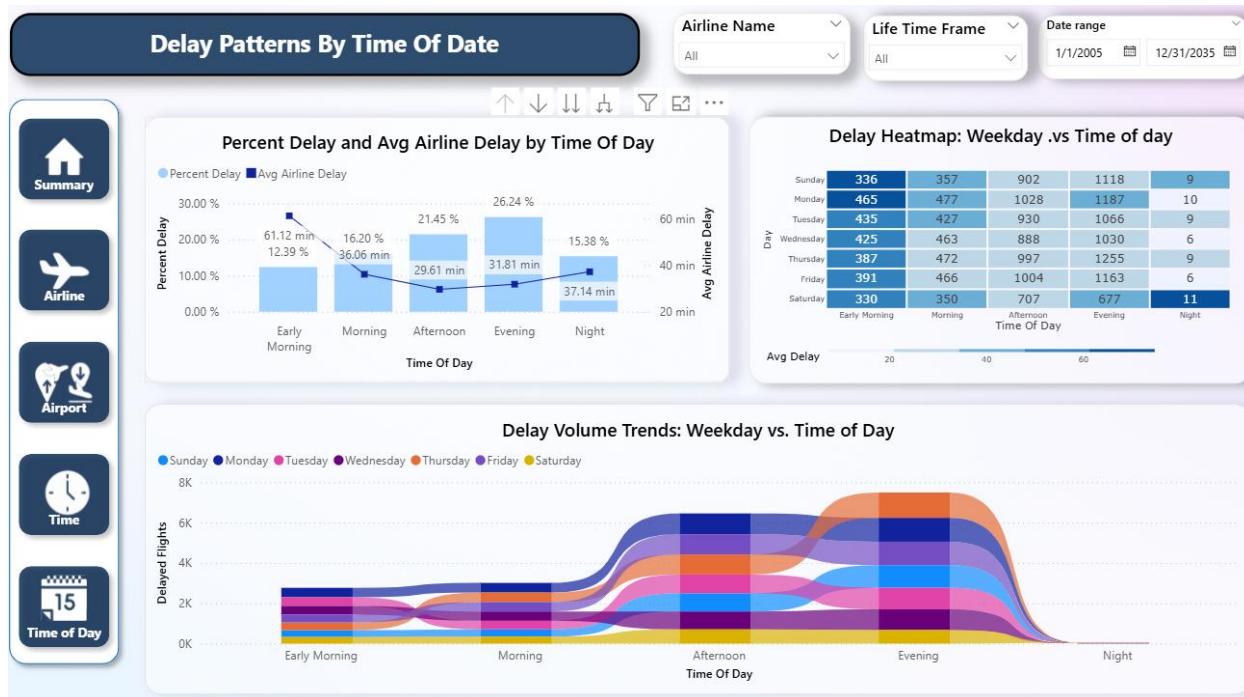
**Biểu đồ:** Line and clustered column chart: Percent Delay & Avg Airline Delay by Time Of Day

#### Insight đạt được:

- Khung giờ Chiều (Afternoon) và Tối (Evening) ghi nhận tần suất trễ chuyến cao nhất, tăng dần từ 20.38% và đạt đỉnh tại 24.79%. Tuy nhiên, thời gian trễ trung bình trong giai đoạn này lại ở mức thấp nhất (khoảng 29–32 phút). Điều này cho thấy việc trễ chuyến vào cuối ngày thường là hệ quả của hiệu ứng dây chuyền (ripple effect) từ sự tích tụ lịch trình trong ngày, dẫn đến số lượng chuyến bị ảnh hưởng nhiều nhưng thời gian chờ từng chuyến không quá dài.

- Ngược lại, khung giờ Sáng sớm (Early Morning) có tỷ lệ trễ thấp nhất (chỉ 10.91%) nhưng lại đối mặt với thời gian trễ trung bình cao kỷ lục, lên tới 61.12 phút – gấp đôi so với khung giờ Tối. Điều này phản ánh rằng các sự cố xảy ra vào đầu ngày tuy ít nhưng rất nghiêm trọng, thường liên quan đến các vấn đề kỹ thuật tàu bay, chuẩn bị nguồn lực hoặc quy trình khởi động hệ thống chưa trọn tru.
- Kết quả phân tích chỉ ra mối tương quan nghịch đảo giữa tần suất trễ và thời gian trễ (Duration vs. Frequency). Càng về cuối ngày, khả năng bị delay càng cao nhưng thời gian chờ đợi càng ngắn. Ngược lại, rủi ro chờ đợi kéo dài tập trung chủ yếu vào các chuyến bay đầu ngày dù xác suất xảy ra thấp.

### 5.3. Phân tích Delay theo Time of Day & Hour (Heatmap)



**Biểu đồ:** Delay Heatmap – Weekday vs. Time of Day

#### Insight đạt được:

- Dữ liệu cho thấy sự đối lập rõ ràng giữa số lượng và mức độ delay. Khung giờ Chiều và Tối là điểm nóng về số chuyến trễ, với số lượng tăng mạnh từ buổi Sáng và đạt đỉnh vào buổi Tối, đặc biệt vào Thứ Năm (1.177 chuyến) và Thứ Sáu (1.117 chuyến). Tuy nhiên, thời gian trễ trung bình ở các khung giờ này chỉ ở mức thấp–trung bình, cho thấy delay chủ yếu do quá tải vận hành trên diện rộng.
- Ngược lại, Sáng sớm tuy có ít chuyến trễ hơn nhưng lại ghi nhận thời gian trễ trung bình cao nhất, phản ánh các sự cố đầu ngày ít xảy ra nhưng nghiêm trọng và khó khắc phục.

- Theo ngày trong tuần, Thứ Năm và Thứ Sáu chịu áp lực vận hành lớn nhất, trong khi Thứ Bảy có mức delay thấp hơn rõ rệt, đặc biệt buổi Tối chỉ còn 625 chuyến, giảm gần 50% so với Thứ Năm, cho thấy nhu cầu và cường độ khai thác giảm mạnh vào cuối tuần.

#### **5.4. Phân tích xu hướng số chuyến Delay theo thời điểm trong ngày**

**Biểu đồ:** Ribbon Chart – Delay Volume Trends: Weekday vs. Time of Day

**Insight đạt được:**

- Khung giờ Tối (Evening) và Chiều (Afternoon) là hai "điểm nóng" ghi nhận số lượng chuyến bay trễ cao nhất trong ngày. Cụ thể, số lượng chuyến trễ tăng vọt từ mức trung bình khoảng 300-400 chuyến vào buổi Sáng (Morning) lên mức 800-900 chuyến vào buổi Chiều, và đạt đỉnh điểm trên 1.000 chuyến vào buổi Tối (đặc biệt là Thứ Năm và Thứ Sáu với lần lượt 1177 và 1117 chuyến).
- Đáng chú ý, có sự phân hóa rõ rệt giữa các ngày trong tuần. Thứ Năm và Thứ Sáu chịu áp lực vận hành lớn nhất với dải ruy-băng (ribbon) mở rộng tối đa vào khung giờ Tối. Ngược lại, Thứ Bảy là ngày "dễ thở" nhất khi số lượng trễ giảm mạnh vào buổi Tối (chỉ còn 625 chuyến, thấp hơn gần 50% so với Thứ Năm), cho thấy nhu cầu di chuyển hoặc áp lực khai thác giảm đáng kể vào ngày nghỉ cuối tuần.

## **XII. Data Mining**

### **1. Tổng quan bài toán**

#### **1.1. Mục tiêu bài toán**

- Dự án xây dựng một mô hình học máy có khả năng dự báo xác suất trễ chuyến của các chuyến bay dựa trên các yếu tố lịch trình (thời gian, hãng hàng không, địa điểm sân bay, ...).
- Dữ liệu được lấy từ DDS\_DATH, bao gồm 100000 chuyến bay với 11 thuộc tính.
- Tuy nhiên, cần loại bỏ chuyến bay bị hủy (Is\_Cancelled = 1) bởi vì khi các chuyến bay bị huỷ, giá trị Is OTP và Is Delay đều NULL, không có ý nghĩa trong việc xác định chuyến bay bị trễ hay không.

#### **1.2. Định nghĩa “chuyến bay trễ”**

“Chuyến bay trễ” là chuyến bay có độ trễ hơn 15 phút và không bị huỷ.

### **2. Khám phá dữ liệu (Exploratory Data Analysis - EDA)**

#### **2.1. Bài toán Mất cân bằng (Class Imbalance)**

- Thực trạng: tỷ lệ Đúng giờ/ Trễ là ~4.4:1
- Vấn đề: Trong dự báo rủi ro (như trễ chuyến, nợ xấu hay bệnh tật), dữ liệu luôn mất cân bằng. Nếu ta dùng thuật toán thông thường, mô hình sẽ “lười

biếng” và dự đoán mọi chuyến bay đều đúng giờ để đạt độ chính xác là hơn 80%. Điều này là vô giá trị đối với vấn đề thực tế.

- Giải pháp: Chọn thuật toán có khả năng xử lý imbalance tốt (sử dụng scale\_pos\_weight trong XGBoost).

## 2.2. Phân tích correlation matrix

Các features tương quan cao với delayed:

- Airline\_Hist\_Delay - Tỷ lệ trễ lịch sử của hãng hàng không
- Airport\_Hist\_Delay - Tỷ lệ trễ lịch sử của sân bay
- Hour - Giờ bay (giờ cao điểm trễ nhiều hơn)
- Is\_Peak\_Hour - Cờ giờ cao điểm

## 2.3. Feature Engineering

### 2.3.1. Chiến thuật chống rò rỉ dữ liệu

- Sắp xếp thời gian: Toàn bộ dữ liệu được sắp xếp nghiêm ngặt theo Date\_Key và TimeOfDay\_Key.
- Kỹ thuật Shift & Rolling: Khi tạo đặc trưng lịch sử (Airline\_Hist\_Delay), sử dụng hàm .shift(). Điều này đảm bảo tại thời điểm dự báo cho chuyến bay T, mô hình chỉ được nhìn thấy dữ liệu của các chuyến bay từ T-1 trở về trước.

### 2.3.2. Tạo đặc trưng mang tính nghiệp vụ

- Route features (Tuyến bay)
  - Kết hợp sân bay đi (Origin\_Airport\_Key) + sân bay đến (Destination\_Airport\_Key) thành 1 feature.
  - Mục đích: Capture đặc điểm của từng tuyến bay cụ thể
- Congestion features (Độ tắc nghẽn sân bay)
  - Đếm số chuyến bay trong 20 observations gần nhất tại mỗi sân bay.
  - Mục đích: Đo lường mức độ tắc nghẽn tại sân bay
- Historical delay features (Tỷ lệ delay lịch sử)
  - Dùng .shift() để tránh data leakage
  - Mục đích: xem hãng hàng không và sân bay này gần đây có delay nhiều không.
- Time features
  - Xử lý tính cyclical của thời gian
  - Practical: Dễ dàng cho model học
- Is\_Peak\_Hour
  - Mục đích: đánh dấu các chuyến bay trong giờ cao điểm

### 3. Thuật toán XGBoost

#### 3.1. Giới thiệu thuật toán

- XGBoost (eXtreme Gradient Boosting) là một phương pháp triển khai nâng cao của thuật toán tăng cường độ dốc. Tăng cường độ dốc là một kỹ thuật học máy mà ý tưởng chính là kết hợp nhiều mô hình đơn giản, còn được gọi là “mô hình học yếu”, để tạo ra một mô hình tổng hợp có khả năng dự đoán tốt hơn.
- Giống như các phương pháp boosting khác, là thêm các mô hình mới vào tập hợp một cách tuần tự. Tuy nhiên, không giống như các phương pháp bagging như Random Forest, nơi các cây được phát triển song song, các phương pháp boosting huấn luyện các mô hình lần lượt từng cái một, mỗi cây mới giúp sửa chữa các lỗi do cây đã được huấn luyện trước đó gây ra.

#### 3.2. So sánh các thuật toán

Thuật toán	Decision Tree	Random Forest	XGBoost
<b>Khái niệm</b>	Là một loại thuật toán máy học đưa ra quyết định bằng cách đặt ra một loạt câu hỏi.	Thay vì chỉ có một cây quyết định đưa ra tất cả các quyết định, chúng ta tạo ra cả một “rừng” cây quyết định. Mỗi cây đưa ra “ý kiến” hoặc dự đoán của mình dựa trên dữ liệu mà nó đã thấy. Kết quả cuối cùng được xác định bằng cách xem xét kết quả đầu ra của tất cả các cây trong rừng.	Huấn luyện các mô hình lần lượt từng cái một, mỗi cây mới giúp sửa chữa các lỗi do cây đã được huấn luyện trước đó gây ra.
<b>Ưu điểm</b>	<ul style="list-style-type: none"><li>- Tính đơn giản: Cây quyết định trực quan và dễ hiểu.</li><li>- Dễ giải thích "tại sao chuyến bay này bị trễ"</li></ul>	<ul style="list-style-type: none"><li>- Robust hơn với noise</li><li>- Có thể xử lý imbalance bằng class_weight='balanced'</li><li>- Feature importance tốt</li><li>- Có thể song song hóa: việc huấn luyện các cây có thể thực hiện song song → thời gian rút ngắn</li></ul>	<ul style="list-style-type: none"><li>- Xử lý imbalance hoàn hảo.</li><li>- Tối ưu AUC thay vì accuracy.</li><li>- Tốc độ: sử dụng nhiều lõi trên CPU để huấn luyện</li></ul>

			mô hình nhanh hơn. - Tính ổn định: tích hợp tham số điều chỉnh giúp tránh hiện tượng quá khóp.
Nhược điểm	<ul style="list-style-type: none"> <li>- Overfitting nghiêm trọng với 98k samples</li> <li>- Không xử lý được class imbalance</li> <li>- Performance thấp</li> </ul>	<ul style="list-style-type: none"> <li>- Vẫn có thể bias về majority class</li> <li>- Memory usage cao với hơn 100 trees</li> <li>- Khó tối ưu cho imbalance nặng</li> <li>- Thời gian dự đoán lâu hơn</li> </ul>	<ul style="list-style-type: none"> <li>- Dễ bị quá khóp nếu không được tinh chỉnh đúng cách</li> <li>- Khó diễn giải hơn</li> </ul>
Ứng dụng	<p>Một số lĩnh vực như:</p> <ul style="list-style-type: none"> <li>- Chăm sóc sức khỏe</li> <li>- Tài chính</li> <li>- Marketing</li> </ul>	<p>Ứng dụng rộng rãi trong nhiều ngành công nghiệp:</p> <ul style="list-style-type: none"> <li>- Ngành ngân hàng: dự đoán khả năng vỡ nợ → giúp quản lý rủi ro</li> <li>- Thương mại điện tử.</li> <li>- Y tế: dự đoán nhiều loại bệnh khác nhau dựa trên tiền sử bệnh án</li> <li>- Thị trường chứng khoán.</li> <li>- Hệ thống đề xuất</li> </ul>	<p>Sử dụng trong nhiều ứng dụng:</p> <ul style="list-style-type: none"> <li>- Dị thường Phát hiện</li> <li>- <b>Phân tích dự đoán</b></li> <li>- Xử lý ngôn ngữ tự nhiên</li> <li>- Hệ thống đề xuất</li> <li>- Chẩn đoán y khoa</li> </ul>

### 3.3. Lý do chọn XGBoost

- Khả năng giải quyết triệt để sự mất cân bằng dữ liệu (Class Imbalance)
  - Trong bài toán 2015\_Flights\_Delayed\_Cancelled, tỷ lệ trễ chuyến chỉ chiếm 18.5% (tỷ lệ 4.4:1). Các thuật toán thông thường sẽ bị "đánh lừa"

bởi lớp đa số, dẫn đến độ chính xác ảo nhưng lại bỏ sót các ca trễ chuyến thực tế.

- Chiến thuật thực thi: tận dụng tham số scale\_pos\_weight = 4.4, giúp mô hình tập trung "phạt" nặng các sai số trên lớp thiểu số (trễ chuyến).
- Xử lý tối ưu dữ liệu dạng bảng đa dạng: Dữ liệu chuyến bay là sự kết hợp của nhiều loại biến số phức tạp:
  - Biến liên tục (Numerical): Airport\_Congestion, Historical\_Delay.
  - Biến phân loại (Categorical): Airline\_Key, Airport\_Key.
  - Biến tính chu kỳ (Cyclical): Hour\_Sin/Cos.
  - Độ phong phú cao (High Cardinality): Route\_Key với hơn 4,000 giá trị khác nhau.

➔ XGBoost được lựa chọn vì khả năng tương thích linh hoạt và cơ chế vận hành thông minh: Thuật toán cho phép chúng ta chủ động kiểm soát cấu trúc dữ liệu đầu vào thông qua One-hot Encoding (đối với các biến danh mục ít nhóm) và Label Encoding (đối với các biến có độ đa dạng cao như Route\_Key) để tối ưu hóa không gian đặc trưng.

- Sự cân bằng giữa Hiệu năng (Performance) và Khả năng giải thích (Interpretability)
  - Robustness: Thông qua quá trình RandomizedSearchCV, chúng ta đã tìm ra bộ tham số: learning\_rate=0.01 (học chậm để tránh nhiễu), subsample=0.7 và colsample\_bytree=0.6 (chống Overfitting).
  - Explainability: XGBoost cung cấp feature\_importances\_, giúp ta xác định được các yếu tố then chốt gây trễ chuyến như: Tỷ lệ trễ lịch sử của hãng bay (Airline\_Hist\_Delay) và Áp lực sân bay (Airport\_Congestion). Điều này giúp các nhà quản lý đưa ra quyết định can thiệp thực tế.

## 4. Chuẩn bị dữ liệu và Huấn luyện mô hình

### 4.1. Chiến thuật chống rò rỉ dữ liệu (Data leakage)

Đảm bảo tại thời điểm dự báo cho chuyến bay T, mô hình chỉ được nhìn thấy dữ liệu của các chuyến bay từ T-1 trở về trước. Nếu không có bước này, mô hình sẽ đạt độ chính xác ảo cực cao nhưng hoàn toàn thất bại khi triển khai thực tế.

### 4.2. Chiến thuật xử lý mất cân bằng dữ liệu (Class Imbalance)

- Dữ liệu cho thấy tỷ lệ Đúng giờ/Trễ là 4.4 : 1.
- Giải pháp: không cắt xén dữ liệu (Undersampling) vì sẽ làm mất thông tin, cũng không chọn tăng ảo dữ liệu (Oversampling) vì dễ gây nhiễu.
- Cơ chế trọng số: Tham số scale\_pos\_weight  $\approx 4.4$  được tích hợp thẳng vào XGBoost. Điều này buộc mô hình phải tập trung học các đặc điểm của nhóm thiểu số (Trễ chuyến) bằng cách tăng mức phạt khi dự báo sai các ca này.

### 4.3. Tối ưu hóa tham số (Hyperparameter Tuning)

Sử dụng RandomizedSearchCV vì với không gian tham số khổng lồ (hơn 100.000 tổ hợp), GridSearchCV sẽ gây lãng phí tài nguyên và thời gian vô ích. RandomizedSearchCV cho phép kiểm tra các vùng tham số triển vọng nhất một cách thông minh, giúp tìm ra bộ tham số tối ưu (như max\_depth=6, learning\_rate=0.05) trong thời gian ngắn hơn gấp nhiều lần mà vẫn đảm bảo hiệu suất.

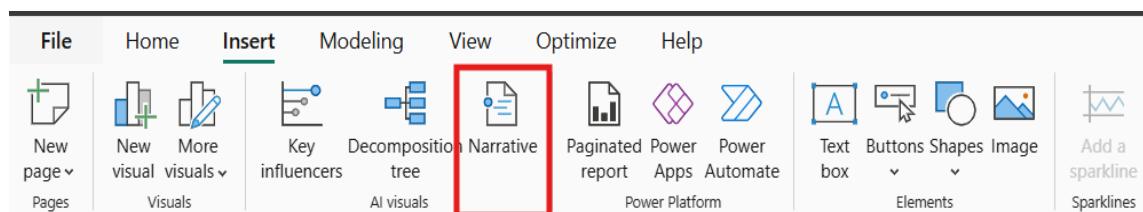
## 5. Đánh giá mô hình

- Không sử dụng Accuracy mà sử dụng AUC để đánh giá, vì trong dữ liệu mất cân bằng, kể cả khi Accuracy 80% cũng có thể là một mô hình "vô dụng" (chỉ đoán đúng lớp đa số).
- Kết quả cho thấy được các yếu tố quan trọng dẫn đến trễ chuyến, 3 yếu tố hàng đầu quyết định trễ chuyến là:
  - Airline\_Hist\_Delay (Lịch sử trễ của hãng): khẳng định tính ổn định trong vận hành của từng hãng.
  - Hour\_Sin (Thời điểm ban ngày/ ban đêm): giá trị là âm → thể hiện ban đêm ảnh hưởng đến delay; giá trị là dương → thể hiện ban ngày ảnh hưởng đến delay.
  - Hour (Thời điểm trong ngày): thể hiện giờ cụ thể trong ngày có khả năng delay.
- Qua đó, ta thấy rằng trễ chuyến không phải là ngẫu nhiên, nó có tính hệ thống. Các hãng hàng không cần tập trung tối ưu hóa lịch trình vào các khung giờ cao điểm (15h-20h) và có phương án dự phòng đặc biệt cho các tuyến bay (Route) thường xuyên gặp sự cố để cải thiện khả năng đúng giờ của chuyến bay.

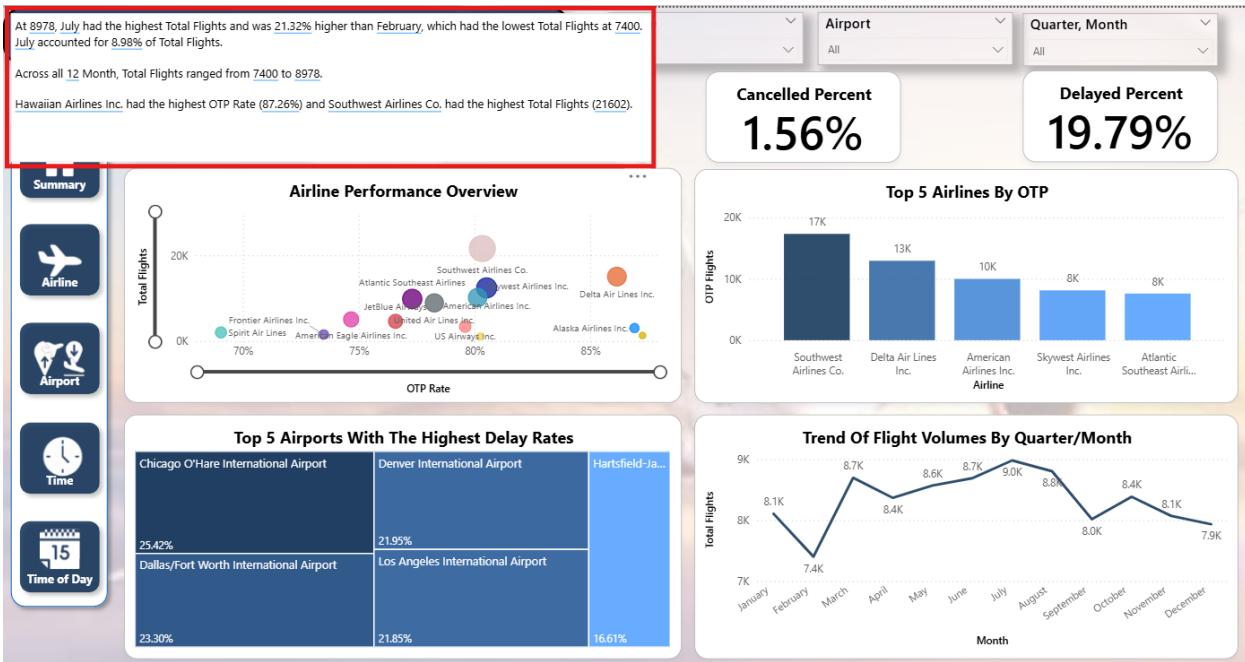
## XIII. Áp dụng AI

### 1. PowerBI Smart Narrative

- Kích hoạt Smart Narrative: Insert -> Narrative.



Sau khi kích hoạt Smart Narrative sẽ cho ra đoạn nhận xét như hình bên dưới.



Narrative tự động cung cấp một cái nhìn tổng quan rõ ràng và có tính diễn giải về dữ liệu khai thác chuyến bay theo tháng và theo hãng.

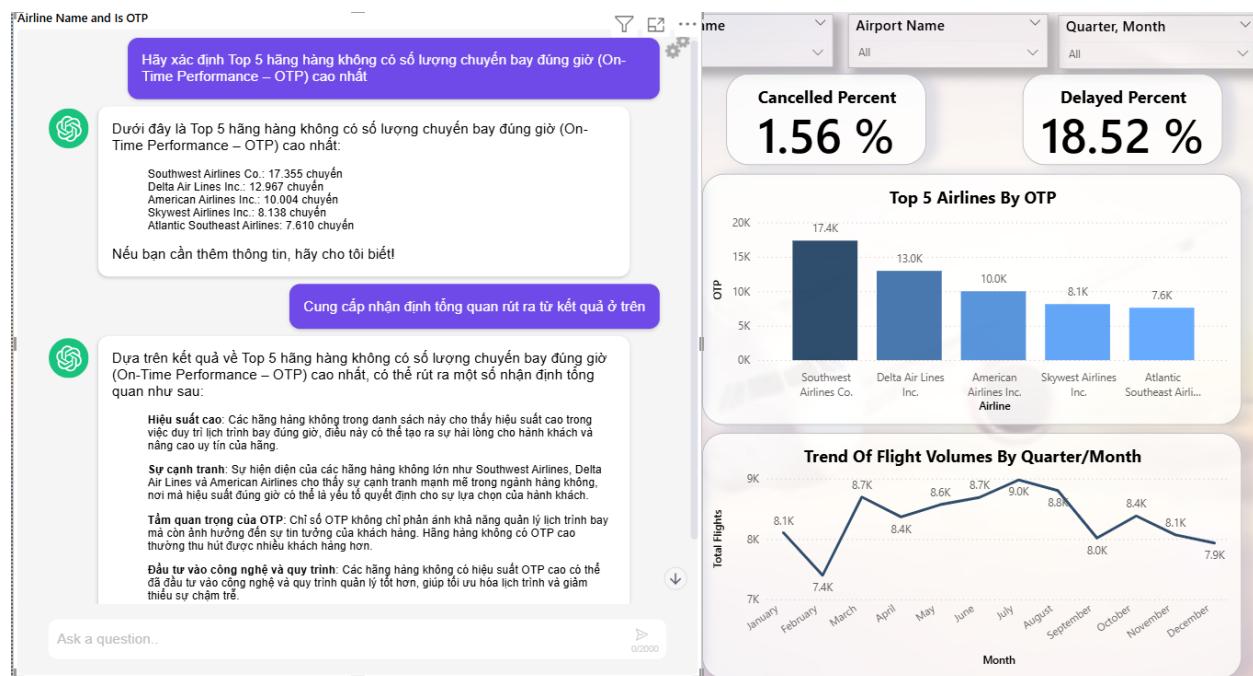
- So sánh giá trị cao nhất – thấp nhất
  - Narrative đã xác định chính xác tháng có tổng chuyến bay cao nhất là tháng 7 (8,978) và tháng thấp nhất là tháng 2 (7,400).
  - Việc chỉ ra chênh lệch 21.32% giúp lãnh đạo dễ dàng nhận diện biến động theo mùa thay vì chỉ nhìn con số tuyệt đối.
- Tỷ trọng đóng góp
  - Việc nêu rõ tháng 7 chiếm 8.98% tổng chuyến bay là rất hữu ích trong việc đánh giá mức độ đóng góp tương đối giữa các tháng trong năm.
- Khoảng giá trị tổng thể
  - Narrative đưa ra khoảng dao động tổng chuyến bay từ 7,400 đến 8,978 trên 12 tháng, giúp xác định mức biến thiên chung và kiểm soát hiệu suất dài hạn.
- Nhận diện hãng dẫn đầu
  - Narrative kết nối phân tích theo tháng với phân tích theo hãng:
    - Hawaiian Airlines Inc. có OTP cao nhất (87.26%)
    - Southwest Airlines Co. có tổng chuyến bay cao nhất (21,602)

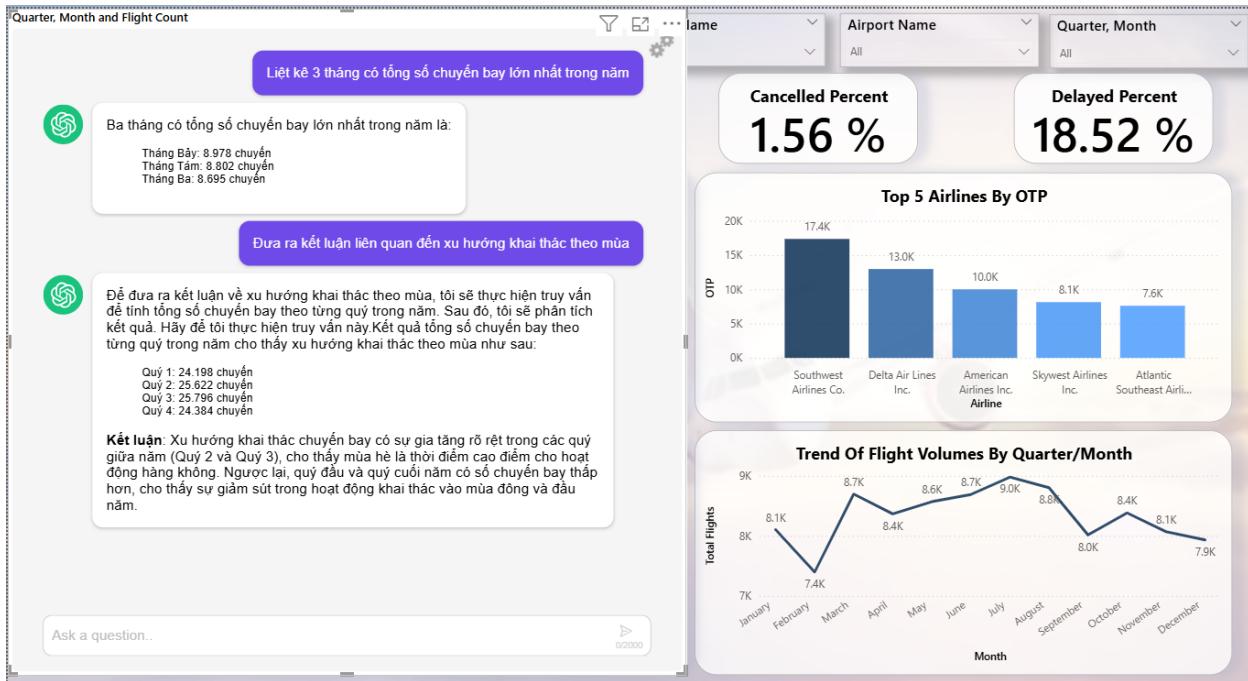
- Điều này giúp người xem hiểu cả hiệu suất chất lượng (OTP) và hiệu suất sản lượng (tổng chuyến bay).

**Kết luận:** Narrative hỗ trợ tóm tắt dữ liệu hiệu quả, cung cấp insight mang tính hành động mà không cần xem trực tiếp biểu đồ. Việc tự động nêu bật các giá trị lớn nhất, nhỏ nhất và mối quan hệ phần trăm giúp rút ngắn thời gian phân tích thủ công và hỗ trợ người ra quyết định tập trung vào các điểm quan trọng nhất của dữ liệu vận hành.

## 2. ChatPowerBI

ChatPowerBI là một tính năng AI tích hợp trong Power BI cho phép người dùng tương tác trực tiếp với dữ liệu bằng ngôn ngữ tự nhiên. Thay vì phải xây dựng biểu đồ hoặc viết biểu thức DAX thủ công, người dùng chỉ cần đặt câu hỏi hoặc yêu cầu bằng tiếng Anh hoặc tiếng Việt và hệ thống sẽ tự động phân tích dữ liệu để trả lời.





ChatPowerBI đã cung cấp các kết quả phân tích chính xác, đầy đủ và bám sát dữ liệu nguồn. Những điểm mạnh thể hiện rõ ràng như sau:

- Khai thác đúng trọng tâm yêu cầu
  - Chatbot trả lời đúng câu hỏi được đưa ra, không lan man
  - Tách biệt nhiệm vụ báo cáo xếp hạng, mô tả và phân tích
- Tự động rút trích insight quan trọng
  - Cho thấy hãng nào đứng đầu về số chuyến bay đúng giờ, và lý do tại sao đáng chú ý
  - Nhận diện tháng cao điểm – thấp điểm và giải thích bằng xu hướng mùa
- Khả năng diễn giải bằng ngôn ngữ tự nhiên
  - Chatbot không chỉ đưa số liệu mà còn diễn giải mối quan hệ và ý nghĩa kinh doanh
  - Giúp người dùng không cần kỹ năng BI hoặc đọc biểu đồ vẫn hiểu được kết quả
- Giá trị hỗ trợ ra quyết định
  - Nhận định về hiệu suất hãng giúp ban điều hành nhắm vào đối tác mạnh/yếu
  - Phân tích theo mùa giúp lập kế hoạch nguồn lực tốt hơn (phi công, tàu bay, lịch bay)

**Kết luận:** ChatPowerBI đã chứng minh khả năng chuyển đổi dữ liệu thô thành thông tin phân tích dễ hiểu và có giá trị thực tiễn. Các kết quả được tạo ra nhất quán với dashboard, chính xác về số liệu và đưa ra insight phù hợp để hỗ trợ nhà quản lý ra quyết định.

## Tham khảo

- [1] *ASPM airport analysis: Definitions of variables*. (n.d.). Faa.gov. Retrieved January 10, 2026, from [https://www.aspm.faa.gov/aspmhelp/index/ASPM\\_Airport\\_Analysis\\_\\_Definitions\\_of\\_Variables.html](https://www.aspm.faa.gov/aspmhelp/index/ASPM_Airport_Analysis__Definitions_of_Variables.html)
- [2] *ASPM: Analysis: Delayed flights*. (n.d.). Faa.gov. Retrieved January 10, 2026, from [http://aspm.faa.gov/aspmhelp/index/ASPM\\_\\_Analysis\\_\\_Delayed\\_Flights.html](http://aspm.faa.gov/aspmhelp/index/ASPM__Analysis__Delayed_Flights.html)
- [3] (N.d.). Bts.gov. Retrieved January 10, 2026, from <http://bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays>
- [4] (N.d.). Bts.gov. Retrieved January 10, 2026, from <http://bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays>
- [5] Wohlwend, B. (2023, July 24). Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning. Medium. Retrieved January 10, 2026, from <https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948>